



JÖNKÖPING UNIVERSITY

School of Engineering

Algorithmic vs. Perceived Fairness in Music Recommender Systems

An Investigation of Popularity Bias from a User Perspective

Main Subject Area: *Informatics*

Specialization in: *User Experience Design*

Author: *Eveline Ingesson*

JÖNKÖPING 2022, June

Certificate of Completion

This final thesis has been carried out at the School of Engineering at Jönköping University within Informatics. The authors are responsible for the presented opinions, conclusions, and results.

Examiner: Bruce Ferwerda

Supervisor: Markus Schedl

Scope: 15 hp (second-cycle education)

Date: 2022-06-07

Attestation of Authorship

I hereby declare that this submission is my work, based upon research that I have conducted.

To the best of my knowledge and belief, it contains no material published or written by another person – except where explicitly defined in the Acknowledgements or listed in the References and properly cited.

Nor does it contain any material of mine that, to a substantial extent, has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Eveline Ingesson

Acknowledgements

This work has been part of a research project conducted by Jönköping University in Sweden and Johannes Kepler University Linz in Austria, and I would, first of all, like to thank both parties for the opportunity to contribute to the project. I have learnt a lot.

I also want to thank my supervisor, and everyone involved in the project for their feedback, valuable input, and the fruitful discussions that have contributed to the outcome of this thesis.

The algorithms used in this study, and the evaluation of them, have been the work of Michaela Berndl, master's student at JKU. Without her work and numerous hours spent helping me implement the algorithms in the platform used for the study, this thesis would not have been possible.

I also want to thank my family for their encouragement and support throughout my entire education. Lastly, I want to thank Ivan, who has been guiding me every step of the way. You are the best.

Abstract

Recommender systems have the potential of helping users in finding relevant items in the online environment, and in many ways, they impact which content we consume. Thus, how fair these systems are affects us. A common fairness issue in recommender systems is popularity bias. However, research on this issue has mostly been focusing on the algorithmic side, and the user perspective has been more or less neglected.

In this study, the goal was to understand whether there is a correlation between algorithmic and perceived fairness in the context of popularity bias, and the study was conducted in a music recommender setting. Three different algorithms were used in the study, each generating recommended playlists with varying levels of fairness in terms of recommending both popular and less popular music items. By comparing how fair users perceived the different recommended playlists to be with the algorithmic fairness of the playlists, conclusions could be drawn on the relationship between perceived and algorithmic fairness. Moreover, it was examined whether two different factors, namely familiarity and satisfaction, have an impact on perceived fairness.

An online survey was conducted, and it was concluded that there is no correlation between perceived and algorithmic fairness, as the participants could not notice any difference in fairness between the playlists. Familiarity was shown to only have an impact on perceived fairness in terms of one algorithm, while satisfaction was shown to have a significant impact on perceived fairness across all algorithms. The results indicate that fairness, in the context of popularity bias, may not be of high importance to users. As opposed to concentrating on how users perceive this type of fairness in recommender systems, it might be more important to focus on other stakeholders, such as the providers.

Keywords: Recommender Systems, Perceived Fairness, Algorithmic Fairness, Popularity Bias, Satisfaction, Familiarity

Table of Content

Certificate of Completion	i
Attestation of Authorship.....	ii
Acknowledgements.....	ii
Abstract	i
Table of Content.....	ii
I Introduction	5
1.1 PROBLEM STATEMENT	7
1.2 PURPOSE AND RESEARCH QUESTIONS.....	8
1.3 SCOPE AND DELIMITATIONS.....	9
1.4 OUTLINE	9
2 Theoretical Framework	11
2.1 FAIRNESS IN RECOMMENDER SYSTEMS.....	11
2.2 POPULARITY BIAS	12
2.2.1 The Impact of Popularity Bias	12
2.3 OPINIONS AND PERSPECTIVES ON FAIRNESS	14
2.3.1 The Impact of Satisfaction	14
3 Method and Implementation	16
3.1 METHOD CHOICE: SURVEY	16
3.1.1 Online Survey Platform	16
3.2 DATA COLLECTION	18
3.2.1 The Structure of the Survey	18
3.2.2 The Design of the Questions	20
3.2.3 The Algorithms	21
3.2.4 Recruitment.....	24
3.3 METHODS FOR DATA ANALYSIS	24

3.4	VALIDITY AND RELIABILITY	26
3.5	CONSIDERATIONS.....	27
4	Results.....	28
4.1	COLLECTED DATA	28
4.1.1	Demographics	28
4.1.2	Perceived Fairness, Satisfaction, and Familiarity	29
4.2	DATA ANALYSIS	30
4.2.1	The Correlation between Algorithmic and Perceived Fairness	30
4.2.2	The Effect of Familiarity on Perceived Fairness	31
4.2.3	The Effect of Satisfaction on Perceived Fairness	32
5	Discussion	34
5.1	RESULTS DISCUSSION	34
5.1.1	Algorithmic vs. Perceived Fairness	34
5.1.2	Familiarity and Perceived Fairness	35
5.1.3	Satisfaction and Perceived Fairness	36
5.2	METHOD DISCUSSION	36
6	Conclusions and Further Research	38
6.1	CONCLUSIONS	38
6.1.1	Practical implications	39
6.1.2	Scientific implications.....	39
6.2	FURTHER RESEARCH.....	40
7	References	41
8	Appendices.....	46
8.1	APPENDIX A: SURVEY PLATFORM.....	47
8.2	APPENDIX B: QUESTIONNAIRE	52
8.3	APPENDIX C: QUESTIONNAIRE RESPONSES	53

8.4	APPENDIX D: FAMILIARITY RESPONSES.....	56
-----	--	----

1 Introduction

Recommender systems are used to assist users in finding relevant items that could be of interest to them (Abdollahpouri et al., 2019). They are algorithmic tools, and they are deployed across many different types of online platforms. The development of recommender systems began in the 1980s (Kunaver & Požrl, 2017), and today, many of us interact with them on a daily basis. We encounter them, for example, when receiving suggestions on products that we might like on e-commerce websites, and when being recommended new connections on social media platforms.

Music recommender systems are a type of recommender systems that we come across on music streaming platforms. These systems aim to provide listeners with recommendations that cater to their preferences and needs (Bauer et al., 2017). In other words, the goal of music recommender systems is to recommend music items that are satisfactory to users (Melchiorre et al., 2021). Music recommender systems provide us with suggestions on, for example, artists, tracks, and albums that we might be interested in.

The performance of a recommender system is typically measured in terms of accuracy (Deldjoo et al., 2019), which is defined by how well the users' predicted preference match their actual preference (Kim et al., 2021). In more technical terms, Koutsopoulos and Halkidi (2018) describe accuracy as "the mean squared error between predicted and true ratings in the training dataset". However, according to McNee et al. (2006), a high level of accuracy does not mean that the recommendations are useful and satisfactory to users. In their paper, they suggest other evaluation criteria for recommender systems that are more centered around the users – similarity, serendipity, and user experiences and expectations. In a different paper by Steck (2011), serendipity is also mentioned as an evaluation criterion, as well as diversity.

Further on, McNee et al. (2006) state that we need to take a more user-centric approach when it comes to evaluating recommender systems. According to the authors, users are not interested in which algorithm has better scores when it comes to different criteria. Instead, what the users want are good recommendations. Thus, as the goal of recommender systems is to help users, it is important to understand their needs as well as their perspective.

Some criteria can be evaluated algorithmically as well as from a user perspective. One of these criteria is fairness, a concept within recommender systems that has been gaining more and more attention recently (Shrestha & Yang, 2019). The consequences of unfair recommender systems are many. For example, unfair recommenders prevent users from finding the items that they want and need, they affect the livelihood of those providing the items, and they can result in disadvantages for minority groups (Deldjoo et al., 2021). Thus, unfair recommenders can have a negative impact on people's lives,

and this might be the reason behind the topic becoming increasingly prevalent in research.

As recommender systems are part of our everyday lives, it is important to make sure that users, as well other stakeholders, are treated equally fair by the systems. This can be done by improving the algorithmic fairness. How users perceive fairness in recommender systems also needs to be considered. If we can understand how users perceive fairness, we can develop better recommender systems that cater to the wants and needs of the users. After all, the goal of using recommender systems is to assist users and improve their experiences of a system or platform.

One aspect that leads to unfairness in recommender systems is popularity bias – a phenomenon where the most popular items receive more and more exposure, while less popular items receive less and less exposure (Bauer et al., 2017). The reason why popularity bias exists is that popular items generally have much more rating data than less popular ones. The reason behind this is that recommender systems are trained on user preferences, and typically, many users rate the popular items, while the less popular items receive only a few ratings. This means that the unpopular items, which constitute the so-called long tail of recommendations, do not get exposure, especially when they are new to the system (Yalcin & Bilge, 2021).

This imbalance in rating data affects the algorithms, which in turn recommend the popular items more frequently than the less popular ones. The more rating data the items have, the more they get recommended – and the more the items get recommended, the more ratings they get from users. This turns into a loop that goes on and on, and the imbalance in recommendations continually grows larger. The rich get richer, and the poor get poorer (Abdollahpouri et al., 2019).

According to Elahi, Abdollahpouri, et al. (2021), most research related to fairness in recommender systems has focused on fairness from an algorithmic viewpoint, and there is a plethora of studies which have introduced methods for improving the algorithmic fairness in recommender systems. This objective approach, however, does not consider the user experience, and how users perceive fairness. Even though recommendations might be fair from an algorithmic standpoint, it does not mean that the recommender system is perceived as being fair by its users (Elahi, Abdollahpouri, et al., 2021). To the best of my knowledge, there is a research gap regarding users' perceptions of fairness, or the lack of fairness, in recommender systems, and it is an area that needs to be explored. This is recognized by Shin and Park (2019), who state that individuals' perceptions of fairness is an important topic for future research.

The purpose of this thesis was to investigate popularity bias from a user perspective, and to try to understand how users perceive popularity bias in the context of music recommender systems. It made use of different algorithms to test whether algorithmic

and perceived fairness correlate, that is, whether recommendations that are objectively fairer are actually experienced as being fairer, and whether recommendations that are less fair are perceived as being less fair. Moreover, this thesis looked into whether different factors affect how users perceive how fair a recommender system is in terms of recommending music items of varying popularity. The investigated factors were familiarity and satisfaction.

1.1 Problem Statement

During the last decade, there has been a significant increase in implementing and using recommender systems (Deldjoo et al., 2021), with the goal of assisting users in finding relevant information. The information that we find impact what content we are exposed to, the decisions that we make, and this shapes us as individuals. In the context of music, recommender systems have a great impact on what content users consume, as streaming platforms are one of the primary sources of music consumption as of today (Ferraro et al., 2021b).

Popularity bias is one aspect of unfairness that has become a widely recognized issue in recommender systems. As it prevents various entities from having a fair chance of exposure and representation, it is an issue of social justice (Abdollahpouri, 2019). Moreover, popularity bias hinders the opportunity for users to discover a greater variety of items. Specifically, it hinders users from discovering items that are not very popular, even though they are good matches for them. The result of popularity bias is that less popular items, which may be of higher quality than the recommended popular ones, never reach the users.

According to Abdollahpouri et al. (2019), popularity bias makes the market more homogeneous, as it becomes dominated by only a few item providers, which in turn leads to “fewer opportunities for innovation and creativity”. Thus, it is important to study this phenomenon, not only from a technical viewpoint, but from different angles and perspectives – such as the perspectives of users’ and other stakeholders. Additionally, popularity bias does not only have a negative impact on the users of a recommender system, but also on the providers of the recommended items as well as the system itself. When the providers’ items do not get recommended, their livelihood is jeopardized. This means that the system fails to keep the providers satisfied, and if the providers are unsatisfied, they may stop using the system. It is due to its detrimental effects on all of the different stakeholders of recommender systems that popularity bias is the fairness aspect that was chosen for this thesis.

As users are important stakeholders when it comes to recommender systems, it is crucial that we understand their perceptions and take their experiences into account when developing recommender systems. Even if we create objectively fair recommender systems that mitigate popularity bias, the users’ perception of how fair they are may or

may not be affected. If we fail to understand the relationship between algorithmic and perceived fairness, and the factors that contribute to a specific perception, we remain limited in creating enhanced user experiences. If we, on the other hand, understand this relationship, we can create recommender systems that cater to the needs of different types of users and meet their expectations. Ultimately, this means that we can develop better applications and platforms.

As previously described, this thesis aimed to understand how users perceive unfairness in terms of popularity bias, and it investigated whether different variables, namely familiarity and satisfaction, affect this perception. There is a research gap when it comes to understanding how users perceive fairness and whether this perception correlates with algorithmic fairness, as most research related to fairness in recommender systems has focused on the algorithmic perspective (Elahi, Abdollahpouri, et al., 2021). In other words, the main focus has, so far, been on algorithmic fairness. This thesis went beyond the objective, algorithmic fairness and took on the user perspective instead.

This study can be one of the starting points in the area of perceived fairness, which future research can build upon. Popularity bias is, in general, a common issue in recommender systems (Kowald et al., 2020). It is a topic that is not only relevant to the research community, but it can also be of interest to various companies and organizations in the industry who wish to improve the experience of their users. Despite the fact that this work was conducted in the context of music, it might also be of relevance to other recommender domains.

1.2 Purpose and Research Questions

Following from the problem statement, we can draw the conclusion that there is a need for research on the topic of perceived fairness. A recommender system may be algorithmically fair, but this does not mean that users perceive it as being fair. Thus, the purpose of this thesis was to understand the relationship between algorithmic and perceived fairness, and it did this by comparing users' perceptions of how fair different recommended music playlists were to the algorithmic fairness of the playlists. Secondly, it investigated whether familiarity and satisfaction have an impact on this perception.

The first research question aimed to answer whether there is a correlation between algorithmic and perceived fairness in terms of popularity bias. Hence, the study's first research question was:

RQ 1: Does perceived fairness correlate with algorithmic fairness in recommended playlists?

The second research question focused on familiarity and its possible implications on perceived fairness. Thus, the study's second research question was:

RQ 2: Does familiarity with the recommended music items in a playlist affect how users perceive fairness?

The third and final research question aimed to answer whether satisfaction affects perceived fairness. Consequently, the study's third research question was:

RQ 3: Does satisfaction with a recommended playlist affect how users perceive fairness?

1.3 Scope and Delimitations

This study was an investigation of popularity bias from a user perspective, with the goal of understanding the relationship between algorithmic and perceived fairness. It explored whether there is a correlation between algorithmic fairness and perceived fairness, that is, whether recommended lists that are fairer in terms of including both popular and unpopular music items are perceived as being fairer and vice versa. This work was also concerned with understanding whether satisfaction and familiarity have an impact on perceived fairness.

This thesis does not provide any in-depth analysis of the algorithms deployed in the study. The focus of this work lies merely on the users' perceptions – not on the algorithms used for data collection. Moreover, this study only examined two different factors – familiarity and satisfaction – and whether they affect perceived fairness. There are numerous factors and variables that could be considered, such as personality traits, musical expertise, music preferences, and demographics. Nevertheless, any evaluation and analysis of additional factors were outside the scope of this research.

Fairness is an extensive area, and there are multiple perspectives that are highly relevant to research. This thesis, however, was limited to looking at popularity bias, which is only one aspect of fairness. Other types of fairness perspectives were not examined. Further on, this research was conducted in the context of music, using music recommender systems. Although the study may be of interest to other types of recommender systems, it cannot be concluded that this work is applicable to them.

In this study, focus lies on perceived fairness from a user perspective. There are other stakeholders whose perceptions could be examined, such as the providers of the recommended items. However, in this thesis, the users of recommender systems were the only stakeholders that were considered.

1.4 Outline

The following chapter presents existing research related to this study. It starts with an overview of fairness in recommender systems and a discussion on popularity bias. Subsequently, research related to the user perspective on fairness is presented.

Chapter 3 describes the methods used in the study, the platform used to collect data, as well as the actual data collection. The chapter ends with a discussion on the methods used for data analysis, the validity and reliability of the thesis, and considerations made when designing the study.

The subsequent part of the thesis, Chapter 4, presents the collected data along with the outcome of the analysis. Chapter 5 starts with a discussion of the results in terms of each of the research questions. In the second section of the chapter, the strengths and weaknesses of the methods used in the study are discussed. Lastly, in Chapter 6, conclusions are presented and suggestions for further research are presented. The report ends with a list of references and appendices.

2 Theoretical Framework

This chapter discusses fairness in recommender systems, popularity bias, the user perspective on fairness, as well as factors influencing opinions on fairness. It does not only put forward studies that are related to the topic explored in this thesis, but it also explains concepts that are of importance to this work.

2.1 Fairness in Recommender Systems

Recommender systems are artificial intelligence (AI) systems which, based on different factors, present users with recommended items. These factors include, for example, past user behavior, as well as user and item attributes (Stray et al., 2021). According to Abdollahpouri and Burke (2019), most research on recommender systems has focused on personalization. In other words, it has focused on how to provide users with recommendations that best match their wants and needs. However, the performance of a recommender system can be measured and evaluated based on numerous attributes. One attribute, which has been gaining an increasing amount of attention of recent, is fairness (Abdollahpouri, Mansoury, et al., 2020). According to Milano et al. (2020), research on ethical issues in recommender systems, such as fairness, is needed, as it is a topic that is relatively new.

There are many definitions of fairness in recommender systems, as well as multiple perspectives that can be considered. Schelenz (2021) provides a broad definition of fairness, stating that “fairness refers to the equal treatment of human beings”. However, the concept of fairness is complex to define (Ekstrand et al., 2018; Farnadi et al., 2018), and no universal definition exists within the relevant literature (Elahi, Jannach, et al., 2021). According to Abdollahpouri, Mansoury, et al. (2020), it is not likely that we will come up with a definition of fairness that fits all kinds of applications, as the definition of fairness in recommendations can differ depending on the domain, the characteristics of the users, as well as the system’s fairness goals. As recommender systems consist of multiple stakeholders, there are many aspects of fairness that can be considered, depending on the perspective one takes.

The majority of the research on algorithmic fairness in recommender systems has been focusing on users (Mehrotra et al., 2018). This is confirmed by Abdollahpouri et al. (2017), who state that the end user of a recommender system is usually the only stakeholder that is considered when evaluating the success of an algorithm. However, users are not the only stakeholders that are affected by fairness issues. Burke (2017) discusses three different categories of stakeholders in the context of recommender systems: consumers, providers, and the system itself.

The consumers are the stakeholders who receive the recommendations, the providers are the ones who supply the recommended items, and the system is the platform itself, which matches consumers with providers (Burke, 2017). Fairness for consumers (C-

fairness), providers (P-fairness), and the system (S-fairness) are perspectives that are all of importance and need to be considered when discussing fairness in recommender systems. However, all different notions of fairness cannot be satisfied at the same time (Pierson, 2017). In a multistakeholder environment, such as a music recommender, there will be a trade-off in fairness between the different stakeholders (Smith et al., 2020). The aspect of fairness that is explored in this thesis is popularity bias, a fairness issue that affects all of the above-mentioned stakeholders.

2.2 Popularity Bias

Recommender systems depend on decisions made by algorithms. It has been demonstrated that this algorithmic decision-making can lead to poor results in numerous ways, either because of problems in the algorithm, manipulation of the system, or data sparsity (Bauer, 2019). Algorithms which are deployed in music recommender systems are susceptible to popularity bias, an issue where popular items are prioritized, while less popular items are more or less neglected.

Abdollahpouri and Mansoury (2020) states that popularity bias is a phenomenon where “popular items are recommended even more frequently than their popularity would warrant”. What causes popularity bias is the fact that popular items have a higher amount of rating data than less popular items, as they are rated more frequently by users.

This lack of balance affects the algorithms, which in turn recommend the popular items more often than the less popular ones. This increases the chance that the popular items will be rated – and as the rating data increases, the popular items get recommended even more frequently (Abdollahpouri et al., 2019). The result is a continuous loop that favors the popular items and makes it harder for less popular items to get exposure. This is commonly referred to as the feedback loop (Mansoury et al., 2020). As all users do not want to be recommended popular items (Abdollahpouri et al., 2019), and as popularity does not guarantee high quality (Ciampaglia et al., 2018), popularity bias is typically viewed as an issue.

Popularity bias results in a limited set of recommendations, which mainly consists of popular items (Bauer, 2019). According to Abdollahpouri et al. (2019), most recommendation algorithms produce recommendation lists that often have close to 100% popular items. As popular items are not necessarily of high quality, over-recommending them is unfair. Ultimately, popularity bias leads to a decreased exposure of other items, even if they are good matches for the users (Chen et al., 2020).

2.2.1 The Impact of Popularity Bias

Research has confirmed that the consequence of popularity bias is an unfair system, in terms of the amount of exposure that is given to items and providers with different levels of popularity. Popularity bias is unfair to, and have negative consequences for,

most stakeholders of a recommender system, with the exception of popular providers who benefit from the gained exposure.

Firstly, popularity bias is unfair to the users, since not all users want to be recommended popular items – many users are interested in long-tail and non-popular items rather than popular ones (Abdollahpouri et al., 2019). According to Celma and Cano (2008), it has been shown that popularity bias decreases user satisfaction and impacts the users' ability to find new, non-obvious recommendations. In line with this, Abdollahpouri et al. (2019) established that many recommendation algorithms give users who are interested in non-popular items almost no such recommendations. In their study, the authors analyzed how popularity bias influences the recommended items, making them deviate from the types of recommendations that the user expects to receive from the system. The MovieLens 1M dataset was used, thus, the study was conducted in the context of movie recommendations. According to the results of the study, many algorithms provide users with recommendations that are heavily focused on popular items, despite the user being interested in non-popular items. In general, popularity bias can have a negative effect on all types of users, as popularity and quality do not necessarily correlate (Ciampaglia et al., 2018).

The above-mentioned research by Abdollahpouri et al. (2019) was later replicated in a study by Kowald et al. (2020). In their reproducibility study, the authors conducted research in the context of music instead of movie recommendation, using the LFM-1b dataset. Some of the main findings were that recommendation algorithms are in favor of popular items in the music domain as well, and that users who are interested in unpopular items receive inferior recommendations compared to users who are interested in popular items.

Popularity bias causes unfairness among the providers as well. The result of popularity bias is that providers of varying levels of popularity get treated differently – the popular ones get a lot of exposure while unpopular get less (Abdollahpouri, Burke, et al., 2020). This leads to an imbalance which has negative effects for the less popular providers, as their revenue, and therefore livelihood, is dependent on being recommended to users (Patro et al., 2020).

Further, popularity bias also affects the system, that is, the platform where the recommender system operates, as it is reliant on both its users and providers in order to function. Keeping both parties satisfied is crucial to the system. Unfairness resulting in dissatisfaction from any of the parties can jeopardize the profit, and ultimately the survival of the platform. Hence, mitigating popularity bias is of interest to everyone that is involved in and affected by the recommendation algorithms.

2.3 Opinions and Perspectives on Fairness

The correlation between perceived and algorithmic fairness, as well as factors influencing perceived fairness in terms of popularity bias, have not yet been researched. However, some studies have investigated users' opinions on fairness. Although opinions on fairness and perceived fairness are two separate matters, it is still of interest to shed light on what users, as well as providers, think about the topic.

In an attempt to understand users' opinions on fairness in recommender systems, Smith et al. (2020) interviewed users about their ideas of fairness in recommendation. In their qualitative study, they showed that the participants prioritized provider fairness over accuracy when realizing how a recommender system could negatively impact the providers. However, participants considered fairness to be more important in specific contexts, such as housing, employment, and finances. It was also noted in the study that the willingness to trade personalization for fairness was influenced by the participants' personal biases.

Sonboli et al. (2021) conducted another interview study where user opinions on fairness were investigated. One of their findings was that the concept of provider fairness was something that very few participants had ever thought about. However, when thinking about the impact of provider fairness, several participants considered it to be of importance. Moreover, a need for transparency in recommendations was expressed by the participants, as they wanted to know the reason why they were recommended certain items.

Not all studies have focused on users' opinions on fairness, but the opinions of providers have also been examined. Ferraro et al. (2021a) conducted interviews with artists to understand their view on fairness in recommendations. What they found was that gender fairness was a big concern among the artists, and the participants expressed that they wanted to see female artists getting more exposure in recommendations, in order to achieve gender balance. In a similar study, Ferraro et al. (2021b) explored which fairness aspects artists consider as being relevant in recommender systems. Through semi-structured interviews, they found that gender balance was important to the artists, which is in line with the previously-mentioned study by Ferraro et al. (2021a). Moreover, they found that the artists considered popularity bias to be an issue. According to the authors, all of the participants agreed that less popular music should be recommended by music platforms – not only the most popular music.

2.3.1 The Impact of Satisfaction

This thesis looks into how two different factors – familiarity and satisfaction – affect how users perceive how fair a recommender system is in the context of popularity bias. To the best of my knowledge, this has not yet been studied. However, there is prior research on how satisfaction affect peoples' opinions on fairness in recommender

systems and algorithmic decision-making, and it has been shown that it is a factor that does affect how fair we consider different algorithms to be.

Through conducting workshops and interviews with people from marginalized populations in the United States, Woodruff et al. (2018) investigated how groups of people who might be negatively affected by algorithmic systems feel about algorithmic unfairness. The study showed that algorithmic unfairness elicited concerns and negative feelings among the participants, which could be interpreted as unfairness resulting in decreased satisfaction. Thus, the way that people are affected by an algorithmic system may have an impact on their perspective on fairness and how satisfied they are with the outcome.

This claim is supported by Wang et al. (2020), who conducted an online experiment to gain insights on how different factors influence opinions on fairness in algorithmic decision-making. The study showed that people consider algorithms to be fairer when the algorithmic outcome is in their favor, thus, when they are satisfied with the recommendations, despite the algorithm being very biased against specific demographic groups.

Based on these results, together with the findings from the study by Woodruff et al. (2018), it could be suggested that satisfaction with an algorithmic outcome can have an effect on how we think about fairness. The more satisfied a person is with an algorithmic outcome, the fairer it is considered to be, and vice versa. Wang et al. (2020) suggests that there is a need to consider this bias, which the authors refer to as an “outcome favorability bias”, when algorithmic fairness is evaluated through feedback from users.

3 Method and Implementation

This chapter outlines and describes the research methodology used in the study, a survey, and discusses why this method was used. It also provides an explanation of the artifact, a survey platform, created to conduct the study. Further, the data collection is discussed, as well as data analysis and the validity and reliability of the study. Lastly, the considerations made when designing the study are presented.

3.1 Method Choice: Survey

The chosen research method for this study was a survey. As defined by Tanner (2002), a survey is “the collection of primary data from all or part of a population, in order to determine the incidence, distribution, and interrelationships of certain variables within the population”. The author states that a survey can involve a variety of techniques for collecting data, such as interviews and observations. For the purpose of this study, however, a questionnaire was used.

The survey was conducted online, using an online survey platform. The choice of conducting the survey online was based on the need for a large number of participants, thus, carrying out the survey in this way saved time as opposed to conducting it in person. Moreover, it facilitated recruitment as it was easier to reach a sufficient number of participants and thereby get higher response rates.

Another research method that could have been used for this study is interviews, as this research aimed to understand opinions and perceptions. However, the research questions in this study were not designed to provide in-depth answers of why the participants have certain perceptions, but they aimed to understand whether algorithmic and perceived fairness correlate, and if a relationship between certain factors exists. Hence, for this study, a qualitative method such as interviews was not appropriate. Instead, a descriptive survey was used, a type of survey that has the purpose of describing a certain phenomenon, as opposed to answering ‘how’ or ‘why’ questions about it (Tanner, 2002).

3.1.1 Online Survey Platform

In order to conduct the survey, an online survey platform needed to be developed. When taking the survey, the participants were presented with song recommendations, based on which they answered various questions. In order to generate these recommendations, data on the participants’ listening history was needed, and this data needed to be fetched from an external source through the platform. Moreover, three different recommender systems that generated the recommendations needed to be integrated with the platform. Due to these requirements, an already existing online survey service could not be used, but a new platform needed to be created. In order to develop this platform, the programming languages HTML, CSS, and PHP were used, and the integrated

recommender systems were programmed using Python. For managing the database and storing the answers from the participants, SQLite3 was used.

In order to generate song recommendations for the participants, based on which they answered the survey questions, data about their music preferences was needed. The solution to this was to fetch the participants' listening history from Last.fm, via the Last.fm API. Last.fm is an online music service which tracks the songs that their users listen to on various music services, such as Spotify, Tidal, Deezer, and iTunes. Thus, through the Last.fm API, the participants' listening history could indirectly be fetched from the music services that they use and have connected to their Last.fm account. Hence, participants were required to have a Last.fm account in order to take part in the survey.

Through the Last.fm API, information about the last 2 000 songs that the participant had listened to was retrieved. Once this data had been fetched, it was processed by the three different algorithms that were used for the study, and recommendations were generated in real time. The recommendations were presented in the form of different playlists – each one generated by a different recommender system – and the participants were asked to answer questions about them. More details on how the survey was structured is explained in section 3.2.1.

A requirement for the participants to take part in the survey was that they needed to have a sufficient listening history on Last.fm – out of the 2 000 fetched songs, at least 50 unique songs needed to exist in the dataset used for training the algorithms. This means that the songs were only counted once, even if they had been played by the participants multiple times. This requirement was essential, as the participants needed to have a sufficient listening history for the algorithms to be able to generate somewhat accurate recommendations. This requirement was checked at the beginning of the survey, and if it was not met, the participants were not able to continue with the survey. Moreover, the participants had to be at least 18 years of age. If they were not, they were prevented from taking part in the survey. Moreover, the platform incorporated functionality which prevented the participants from conducting the survey more than once.

To enhance the user experience during the survey, a progress bar was shown at the top of each page. This let the participants know how far they had progressed in the survey, and how much of the survey that was left. Moreover, the goal was to design an aesthetically pleasing interface that was easy to use, and which looked reliable and trustworthy. The idea was that this would minimize the dropout rate and that it would encourage more people to participate in the survey. Additionally, if the participants happened to close the browser during the survey, they were redirected to the last page they were on once they re-entered the website. This functionality was also implemented in order to prevent a high dropout rate. The platform was also responsive, meaning that

participants could complete the survey on different devices, such as a computer, tablet, or a mobile phone.

Prior to collecting the actual data, a series of test runs were performed, both with researchers within the field of recommender systems and people with no prior knowledge of the study. This was to ensure that everything worked as intended, and that the platform had no bugs or other potential issues. Screenshots of the different parts of the survey platform can be found in Appendix A.

3.2 Data Collection

As described in the previous section, the data from the survey was collected through the use of an online survey platform, developed specifically for this study. This study aimed to answer three different research questions, and the results from the survey were used to answer all of them. Thus, no other type of methodology was needed in order to conduct the research.

In order to answer RQ 1, whether algorithmic fairness correlates with perceived fairness with regards to popularity bias, the participants were shown three different playlists with song recommendations, generated by three different algorithms. Based on these song recommendations, they answered questions on how fair they perceived the playlists to be. The participants received personalized recommendations, which means that the fairness of the lists varied between participants. The algorithmic fairness of each playlist could be compared to the perceived fairness, and based on this, conclusions could be drawn on the relationship between algorithmic and perceived fairness. Moreover, as the different algorithms were already ranked in their level of algorithmic fairness, the perceived fairness could be compared to how mathematically fair the algorithms were.

When it came to answering RQ 2, whether familiarity affects how users perceive fairness in the context of popularity bias, the participants were asked to answer which songs in the playlists that they were familiar with. This information was used to understand if there is a relationship between the level of familiarity and perceived fairness.

RQ 3 aimed to answer whether satisfaction affects how users perceive fairness in the context of popularity bias. By asking the users questions on how satisfied they were with the playlists, conclusions could be drawn on whether there is a relationship between satisfaction and perceived fairness.

3.2.1 The Structure of the Survey

This section details the different parts of the survey and presents motivations behind the design choices. Screenshots showcasing the different parts of the online survey can be found in Appendix A.

Before the participants took part in the survey, they were presented with information on what the study entailed. However, they were not informed that the study was about understanding perceived fairness, as their answers might have become biased if they knew too much about the purpose of the study. The information given to the participants was that the study was used to understand people's opinions on different playlists with personalized song recommendations. A short introduction to the different parts of the study was also provided, along with information on how long it would take them to complete the survey.

Subsequently, information about the need of a Last.fm account was given. The participants were informed that no data other than the listening history would be accessed and used, and that their data would not be used for any purposes that were not related to the study. They were also informed that they could exit or withdraw from the survey at any time. If the participants agreed to the conditions, they needed to provide their username on Last.fm, which was used to fetch their listening history from the Last.fm API. When this was done, they could proceed with the survey.

If the participants' listening history on Last.fm was sufficient, meaning that they had listened to at least 50 unique songs which existed in the dataset used for training the algorithms, they were able to participate in the study. This requirement ensured that the generated recommendations were personalized. Throughout the survey, there were two control questions, located on different pages, where the participants were asked to select a specific answer. These control questions made it possible to detect fake and careless responses, which could then be removed during data analysis.

In the first part of the survey, the participants were asked to provide general information about themselves, namely their age, gender, country of origin, and country of residence. On the following page, they took a personality test, namely the Ten-Item Personality Inventory (Gosling et al., 2003). After completing this test, they moved on to a page where they answered six questions related to their musical expertise. The questions used to assess this type of expertise were adopted from Goldsmiths Musical Sophistication Index (Gold-MSI) developed by Müllensiefen et al. (2014). However, the data collected from these two tests was not analyzed in this study but was collected for any potential future research on whether personality and musical expertise can affect perceived fairness.

After these steps, a page with information about how the following part of the survey would work was displayed to the user. Upon continuation, the three different playlists were shown to the participants, one by one. Every playlist was shown on a separate page. This design choice was based on the idea that the participants might not have thought about the different playlists in-depth if they were displayed on the same page, as this might have led to them thinking that the playlists were supposed to be compared and ranked. This was not wanted in this study, and by displaying the playlists on

different pages, direct comparison could be avoided. The recommended playlists were presented in a randomized order, in order to mitigate that a specific order would have an impact on the results.

Each recommended playlist consisted of 10 songs, and the information that was displayed was the song titles, name of the artists, as well as the music genres. Next to each track, there were different checkboxes where the participants indicated whether they were familiar with the song, if they had listened to the song, and if they considered the song to be popular. The information on whether they were familiar with the song or not could be used to answer RQ 2. The remaining data was not analyzed in this study but can be of use for further research.

Next, the participants answered three questions on how satisfied they were with the playlist, and these answers were used to answer RQ 3. These satisfaction questions were adopted from Graus and Ferwerda (2021). Subsequently, the participants answered three questions related to perceived fairness. These fairness questions were developed specifically for this study and were not derived from any existing questionnaire or previous research. The questions on perceived fairness and satisfaction can be found in Appendix B.

Lastly, one question aiming to understand whether the playlist had a higher or lower ratio of popular songs than the participant usually listened to was posed. The data collected from this question was not analyzed in this study but could be of use to future research. On the final page of the survey, the participants were asked to answer what made them consider a song as being popular or not – the song itself, the artist, or the genre.

3.2.2 The Design of the Questions

The questions on fairness and satisfaction were formulated as statements where the participant had five options to choose from: disagree strongly, disagree a little, neither agree nor disagree, agree a little, or agree strongly. Questions formulated in this way are known as Likert-type items. The Likert scale, developed by Likert (1932), is designed to measure attitudes in a scientific way, and it is one of the most frequently employed research tools (Joshi et al., 2015). As described by Boone and Boone (2012), a Likert scale consists of a number of items that are combined into a score, which can be used to, for example, measure personality traits.

Likert-type items can have a varying number of options to choose from, known as points. According to Krosnick and Fabrigar (1997), the optimal length might be between 4–7 points. Too many points might make the meaning of the options less clear. Too few points, on the other hand, might not give the respondents the ability to express their opinions in a precise way. Likert scales most commonly use five points (Krosnick, 2017).

Likert-type items that use five points have a midpoint, that is, a neutral response option. There are both positive and negative aspects of including this. On the one hand, respondents might select the midpoint option as it does not require much effort and is easier to justify (Krosnick & Fabrigar, 1997). On the contrary, not including a midpoint forces the participants to choose either a positive or negative attitude, even though they might have a neutral opinion (Krosnick, 2017). As this study was conducted online and the participants were anonymous, they did not have to justify their choices. Based on this, the survey Likert-type items in this study had five response options to choose from, thereby a midpoint was included.

Another choice one must take when designing Likert-type questions is whether to label the response options with words or numbers. Verbal labels (using words) might result in ambiguity, as language can be ambiguous, which is something that can be mitigated using numeric values. However, people do not commonly express opinions using numbers, and using verbal labels can help the participants express themselves easier and in a more natural way (Krosnick & Fabrigar, 1997). Further on, several studies show that reliability increases when labeling all the options with words, and that respondents are more satisfied with verbal labels (Krosnick, 2017). Thus, in this study, verbal labels were used to describe the different response options.

According to Peterson (2000), questions should be kept as brief as possible, and they should not be longer than 20 words. All the statements in the questionnaire used in this study were less than 20 words long. Moreover, the questions should not include more than three commas, in order to increase understandability (Payne, 2014). The questions used in this study did not contain any commas at all.

The statements in this study were formulated in a non-leading way, in an attempt to mitigate bias as much as possible. The term ‘fair’ was avoided, as it could be difficult for the participants to interpret. Instead, the word ‘balanced’ was used. Three different statements were used to measure each construct (fairness and satisfaction). By including several questions for each construct, it could be derived, during data analysis, whether the questions had captured the constructs that were intended to be measured. All of the questions employed in this study can be found in Appendix B.

3.2.3 The Algorithms

For this study, three different algorithms were used, all of which had varying levels of fairness in terms of including both popular and less popular items. The motivation behind using three algorithms was, first of all, to not make the survey too long. The risk when including a higher number of generated playlists was that the participants would lose focus and start giving careless responses at the end of the survey.

Secondly, the algorithms needed to process the data and generate recommendations in real time. The higher the number of algorithms, and the more complex they were, the

slower the loading time. If it would take several minutes from the time the participants started the survey and the first questions were displayed, there was a risk that many of the recruited participants would leave the platform thinking that it did not work. By using three relatively fast algorithms that processed the data within less than one minute, this could be avoided.

It could be argued that including a higher number of algorithms would give more fine-tuned data on how users perceive fairness. However, as this was the first study investigating perceived fairness in this context, the most crucial aspect was that the recommenders used were clearly different from each other in their level of fairness. For this study, this criterion was more important than including a high number of algorithms which did not differ much in terms of fairness.

The algorithms were trained using the LFM-2b dataset, which, as of 2022, contains the listening data of over 120,000 Last.fm users. In total, the LFM-2b dataset consists of two billion listening events (Schedl et al., 2022). However, for this study, only a sample of this dataset was used for training the algorithms, namely 180 000 tracks. Out of these 180 000 tracks, information about the music genre was included for 150 000 tracks.

The least fair algorithm used in this study was the variational autoencoder (MultiVAE). MultiVAE is a latent variable model that learns a deep representation from high-dimensional data. Given the user’s interaction vector, this algorithm estimates the probability distribution over all items. MultiVAE employs multinomial likelihood and a different regularization procedure involving linear annealing (Liang et al., 2018).

The algorithm used in this study which was ranked in the middle in terms of fairness was the sparse linear method (SLIM). SLIM generates top-N recommendations by aggregating from user profiles. A sparse aggregation coefficient matrix W is learned from SLIM under the L1 and L2 constraints. The learned item coefficients are then used to predict the recommended items of the user (Ning & Karypis, 2011).

The fairest algorithm utilized in this work was the k-Nearest Neighbors (ItemKNN), which is a memory-based recommendation algorithm and a standard method of collaborative filtering. This approach is based on computing the item-item similarity, meaning that an item is recommended to a user if it is similar to items the user already selected. The ItemKNN algorithm uses statistical measures to compute the item-item similarities (Sarwar et al., 2001).

The measures used to determine the differences in fairness originated from Lesota et al. (2021) but were adapted to measure item fairness instead of user fairness. The statistical measures used were Kendall’s Tau, Kullback–Leibler Divergence, and the Delta Metrics Mean (% Δ Mean). In this context, popularity was defined as the number of distinct users that had listened to a song in the training dataset. This was used to define the popularity distribution over the training dataset and the top 50 recommendations of

the user. The measures were computed for each user and then aggregated by calculating the median to receive a value for each algorithm and metric.

In order to measure the difference between the distributions, the $\% \Delta$ Mean popularity between the training data and the recommended lists were calculated. A positive $\% \Delta$ Mean means that the overall popularity of the songs in the recommended lists are higher than in the training data.

The Kullback-Leibler Divergence and Kendall's Tau were used to compare the entire popularity distribution. For calculating these measures, decile-bins were created, such that the cumulative popularity of all tracks of the collection belonging into one bin constitutes approximately 10% of the total popularity of all tracks of the whole collection.

The Kullback-Leibler Divergence measures the difference between two distributions and increases with every mismatch in the item counts. High values for this measure indicate that the overall popularity distributions of the recommended items are highly different from those of the training dataset.

On the other hand, Kendall's Tau is a distance function that counts the number of pairwise disagreements between two ranking lists. Kendall's Tau shows whether there are common patterns in the two distributions, meaning that the order of the bins in the two distributions is similar. The Kullback-Leibler Divergence reaches its maximum value of 1 when the two distributions are identical from the ranking point of view. For fully reversed rankings, the correlation score is -1.

Kendall's Tau showed negative values for all three algorithms, indicating that the ranking of the bins in the recommendation lists do not reflect the ranking in the training data. For MultiVAE and SLIM, the ranking of the bins in the recommendations is rather reverse than in the training dataset. The ItemKNN recommender demonstrates a lower median Kullback-Leibler Divergence than MultiVAE and SLIM, which means that its output better correlates with training data in terms of popularity distribution. ItemKNN and SLIM shows a lower value in $\% \Delta$ Mean compared to MultiVAE, suggesting that MultiVAE favors more popular items in the recommendations. Overall, the measures indicate that the ItemKNN recommender better approximates the popularity distribution in the training data, which is then followed by SLIM. According to the results, MultiVAE gives the most unfair recommendations in terms of recommending tracks of varying popularity. The differences in fairness between the three different algorithms can be seen in Table 1.

Fairness Measures			
	Kendall's Tau	Kullback-Leibler Divergence	Delta Metrics Mean
MultiVAE	-0.66	7.36	1322.01
SLIM	-0.61	4.88	478.06
ItemKNN	-0.0827	1.41	199.69

Table 1: Measures used to determine the differences in fairness between the algorithms.

3.2.4 Recruitment

Participants were recruited from Amazon Mechanical Turk (MTurk), and they were paid for their participation. According to Roscoe (1975), it is appropriate for most research to have a sample size that is larger than 30 and smaller than 500. For this study, the goal was to get at least 100 valid responses.

A requirement for participation in the study was that the participants needed to have a Last.fm account, so that data on their previous listening history could be retrieved. Without access to the listening history, no song recommendations could be made, thus, the survey could not be completed. The participants also needed to be at least 18 years of age in order to participate. There were no other requirements for participation, as there were no specific population that this research aimed to study.

3.3 Methods for Data Analysis

As mentioned in previous sections, data was collected through an online survey platform developed specifically for this study. A total of 170 survey responses were collected, and this data was initially stored in an SQLite3 database. However, all of the responses could not be used for analysis. The responses that were incomplete were filtered out, along with responses from participants who failed to answer the two control questions correctly. After filtering the data, 115 valid responses remained in the dataset, which could be used for analysis.

When the initial processing of the data was completed, the database was converted into a CSV file and exported to the statistical analysis software SPSS for analysis. As the data collected in this study was quantitative in nature, a quantitative data analysis needed to be conducted. Thus, SPSS was deemed an appropriate tool.

In the demographics section of the survey, participants provided information about their age, gender, country of origin, as well as their country of residence. This data was analyzed using descriptive statistics. During analysis, the ages were grouped into categories to provide a clearer overview of the age distribution.

In order to answer RQ 1, which was concerned with whether there is a correlation between perceived and algorithmic fairness in the context of popularity bias, two different types of analysis were used. The first analysis used was Pearson's correlation coefficient, which is a statistical test that measures the extent to which two variables are correlated (Sharma, 2012). As the participants received personalized recommendations, different lists were produced for the different participants. This means that the fairness of the lists varied between participants, even though they were generated by the same algorithms – the participants did not receive playlists with the same level of fairness.

As mentioned in section 3.2.3, three different fairness measures were used to determine the algorithms' level of fairness – Kendall's Tau, the Kullback-Leibler Divergence, and the Delta Metrics Mean. These measures were not only used to rank the algorithms, but they were also used to compute individual "fairness scores" for each participant and algorithm. These computed scores, or values, could then be compared to each participant's perceived fairness value. By using Pearson's correlation coefficient, it could be determined whether a correlation between these values existed or not. Thus, the correlation between algorithmic and perceived fairness for each algorithm could be computed, by considering how fair the recommendations were for each participant.

The second type of analysis used was a repeated measure design, namely repeated measures analysis of variance (ANOVA). This type of analysis is known as a within-subject design, and it was used to determine whether the difference in fairness between the algorithms was perceived by the users.

As described by Lamb (2003), repeated measures ANOVA measures each participant on the same dependent variable several times. In this study, the same participants took part in all of the different conditions in the study, answering questions about all the different recommended playlists. This meant that the participants were measured three times on the same dependent variable, namely perceived fairness. Hence, using repeated measures ANOVA was appropriate in this context.

Through the use of repeated measures ANOVA, the differences in mean scores under the three different conditions could be analyzed. To evaluate the differences in perceived fairness across the different recommended playlists, and to understand whether the mean differences were significant, a pairwise comparison was made. This type of analysis differed from the first one as it compared perceived fairness between the different algorithms rather than analyzing them in isolation.

Prior to conducting the two different analyses, the second and third Likert-type item for measuring perceived fairness were reverse coded. Moreover, for each algorithm, the three questions measuring fairness were computed into one variable, by combining the scores from all the questions and calculating the average of the responses. The same was done for the three satisfaction questions. However, the results proved to be inconsistent and inconclusive when using all of the fairness questions. Therefore, the second and third fairness questions were excluded from analysis, as this gave more sensible and explicable results.

To answer RQ 2, whether familiarity affects how users perceive fairness in the context of popularity bias, one-way analysis of variance (ANOVA) was used. One-way ANOVA is a technique used to understand the differences in a dependent variable based on one independent variable (Allen, 2017). In this case, the dependent variable was perceived fairness, while the independent variable was familiarity.

For answering RQ 3, whether satisfaction affects how users perceive fairness in the context of popularity bias, one-way ANOVA was once again used. In this context, the dependent variable was perceived fairness, while the independent variable was satisfaction.

3.4 Validity and Reliability

Validity refers to the extent to which a study measures what was intended to be measured (Gipps, 2011). Reliability, on the other hand, is defined as the extent to which a study can be replicated (Andres, 2012). The validity and reliability of this study are, in this report, discussed with regards to how the study was designed and evaluated.

Firstly, the survey was conducted online, and not in person, which could be argued to strengthen the reliability of the study. No humans were involved in, for example, explaining the survey or asking questions. The information provided to the participants, along with its delivery, was the same for everyone.

However, the environment could not be controlled, as the participants were free to choose when, where, and on which device they would take the survey. This is a factor that could potentially have an effect on reliability. On the other hand, conducting the survey in a controlled environment would make it difficult to recruit the number of participants needed for the study – which was one of the main reasons for conducting it online in the first place.

Additionally, the satisfaction questions which were used in the questionnaire were adopted from Graus and Ferwerda (2021). In their study, it was shown that these questions measured the same construct. This contributed to an increase in reliability with regard to this part of the questionnaire.

In order to increase the validity of the survey, both the online platform and the questionnaire were reviewed by researchers and professors from Johannes Kepler University Linz in Austria, all of whom are active within the field of recommender systems. During these reviews, the survey platform was tested for potential errors, and the questionnaire was revised to ensure that the questions were not confusing or misleading. Moreover, the survey was pre-tested by people with no prior knowledge of the study to ensure that the information and the questions were clear and understandable.

In order not to bias the participants' responses when taking the survey, no explanations of fairness or popularity bias were provided. Further, the survey included two control questions which helped detect fake and careless responses. After the survey had been conducted, these responses could be removed from the data. Finally, the order of the algorithms was randomized for each participant, which ensured that no specific order had an impact on the results.

3.5 Considerations

Several considerations were addressed when designing the study. Firstly, it was made clear to the participants at the beginning of the study that their data would be handled with care and that it would not be used for anything other than research. Additionally, they were informed that their answers would not be shared with anyone that was not involved in current or future research stemming from the collected data.

The participants were also made aware, before starting the survey, of how it would work, and which information they needed to provide. This gave them the opportunity to not participate if they felt that there was information that they did not wish to share.

In order to participate in the study, the participants needed to provide their username on Last.fm. It was explained to the participants that the listening history was the only information that would be used for the study, and that the fetched data would not be used for any other purposes other than the study. Lastly, it was made clear that participation was completely voluntary, and that the participants could exit or withdraw from the survey at any time.

A requirement for conducting the survey was that the participants had to be 18 years or older. This was an ethical decision, as consent might have been needed from a parent or guardian in order for a minor to participate, which can be problematic when conducting an online survey. If the participants were younger than 18 years, they were not able to take part in the study.

4 Results

In the first section of this chapter, the data collected from the survey is described. In the following section, the analysis is discussed with regards to the research questions that this study aimed to answer.

4.1 Collected Data

Through recruiting on Amazon MTurk, a total of 170 survey responses were collected. After incomplete and faulty responses were filtered out, 115 valid responses remained in the dataset, which could then be used for analysis.

4.1.1 Demographics

Out of the 115 participants, 45 (39.1%) were female and 70 (60.9%) were male, as can be seen in Table 2. Further, the ages of the participants ranged between 21 and 64 years. The distribution of different age groups is demonstrated Table 3. As for the participants' country of origin and country of residence, a clear majority of the participants originated from and resided in the United States of America, followed by India. The distribution of the participants' country of origin and country of residence can be seen in Table 4 and 5.

Gender Distribution					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	45	39.1	39.1	39.1
	Male	70	60.9	60.9	100.0
	Total	115	100.0	100.0	

Table 2: Gender distribution amongst participants.

Distribution of Age Groups					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	21 - 29	34	29.6	29.6	29.6
	30 - 39	44	38.3	38.3	67.8
	40 - 49	26	22.6	22.6	90.4
	50 - 59	8	7.0	7.0	97.4
	60+	3	2.6	2.6	100.0
	Total	115	100.0	100.0	

Table 3: Distribution of age groups amongst participants.

Country of Origin					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Brazil	1	.9	.9	.9
	Colombia	2	1.7	1.7	2.6
	France	1	.9	.9	3.5
	India	28	24.3	24.3	27.8
	Philippines	1	.9	.9	28.7
	Sweden	1	.9	.9	29.6
	United Arab Emirates	1	.9	.9	30.4
	United States of America	80	69.6	69.6	100.0
	Total	115	100.0	100.0	

Table 4: Distribution of the participants' country of origin.

Country of Residence					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Brazil	1	.9	.9	.9
	India	27	23.5	23.5	24.3
	Sweden	1	.9	.9	25.2
	Turkey	1	.9	.9	26.1
	Uganda	1	.9	.9	27.0
	United Arab Emirates	1	.9	.9	27.8
	United Kingdom	1	.9	.9	28.7
	United States of America	82	71.3	71.3	100.0
	Total	115	100.0	100.0	

Table 5: Distribution of the participants' country of residence.

4.1.2 Perceived Fairness, Satisfaction, and Familiarity

For collecting data on perceived fairness and satisfaction, Likert-type items were used in a questionnaire. For each algorithm, the three items measuring satisfaction were computed into one variable. This was done by combining the scores from all the questions and calculating the average of the responses. The frequencies and percentages of the participants' average scores for each algorithm, in terms of perceived fairness and satisfaction, can be found in Appendix C.

Data on the participants' familiarity with each playlist was collected through the participants indicating whether or not they knew the recommended songs. Descriptive

statistics on how many songs the participants were familiar with in each playlist can be found in Appendix D.

At the very end of the survey, participants were asked to indicate which factors influenced them when determining whether a music item was popular or not – the song, the artist, or the genre. The results showed that 27.8% (32 participants) considered the genre when determining whether an item was popular or not, 71.3% (82 participants) considered the artist, and 80.9% (93 participants) considered the song itself.

4.2 Data Analysis

In this section, the data analysis is presented in regard to each of the three research questions, starting with RQ 1, followed by RQ 2, and lastly, RQ 3.

4.2.1 The Correlation between Algorithmic and Perceived Fairness

The first research question aimed to answer whether the perceived fairness of recommended playlists correlates with algorithmic fairness in the context of popularity bias in music recommender systems. In order to analyze the data gathered from the Likert-type items measuring perceived fairness, two methods were used, namely Pearson’s correlation coefficient and repeated measures ANOVA.

Pearson’s correlation coefficient was used to understand the correlation between perceived and algorithmic fairness by looking at each algorithm in isolation. For each user and algorithm, three algorithmic fairness values were computed using Kendall’s Tau, the Kullback-Leibler Divergence, and the Delta Metrics Mean. These values were then compared to the users’ perceived fairness values, in order to see if a correlation existed. In this analysis, the threshold for statistical significance was set to 0.05. The results are shown in Table 6.

Correlations between Perceived and Algorithmic Fairness

		Kendall's Tau	Kullback-Leibler Divergence	Delta Metrics Mean
MultiVAE	Pearson Correlation	-0.76	-0.23	-0.002
	Sig. (2-tailed)	.419	.805	.981
SLIM	Pearson Correlation	-0.87	-.115	-0.99
	Sig. (2-tailed)	.357	.220	.294
ItemKNN	Pearson Correlation	.010	.047	-.064
	Sig. (2-tailed)	.914	.620	.496

Table 6: Correlations between perceived and algorithmic fairness for each of the algorithms. Sig. (2-tailed) represents the p value.

As can be seen in Table 6, all p values exceeded 0.05. This means that there was no significant correlation between the perceived and algorithmic fairness, for any of the algorithms, based on the three different fairness measures.

By using the second analysis method, repeated measures ANOVA, the differences in mean scores for the three different algorithms could be analyzed. This was done in order to understand whether participants perceived a difference in fairness between the algorithms. To evaluate these differences, a pairwise comparison was made using the Bonferroni correction.

When conducting the pairwise comparison between MultiVAE and SLIM, the mean difference was -0.009, and the mean difference between MultiVAE and ItemKNN was -0.078. Further, when comparing SLIM and ItemKNN, the mean difference was -.070. For this analysis, the threshold for statistical significance was set to $\alpha = 0.05$. This meant that if the p value exceeded this, no statistical significance could be shown. In all pairwise comparisons, the p value ($p = 1.000$) exceeded the significance level. This meant that there was no significant difference in perceived fairness across the different algorithms.

4.2.2 The Effect of Familiarity on Perceived Fairness

To answer RQ 2, whether familiarity affects how users perceive fairness in the context of popularity bias, one-way analysis of variance (ANOVA) was used. In this analysis, the dependent variable was perceived fairness, while the independent variable was familiarity. When analyzing the algorithms, the threshold for statistical significance was set to $\alpha = 0.05$.

Firstly, the effect of familiarity on perceived fairness was analyzed for the MultiVAE algorithm. The results showed a positive, F-value, $F(6, 108) = 1.273$, indicating a positive correlation between familiarity and perceived fairness. However, the R-squared value was relatively low ($R^2 = 0.066$), indicating that the relationship was not very strong. Moreover, there was no statistical significance, as the p value exceeded 0.05 ($p = 0.276$). This meant that it was shown that familiarity does not have an impact on perceived fairness when it comes to the MultiVAE algorithm.

Secondly, SLIM was analyzed. In this case, the F-value was also positive, $F(8, 106) = 2.605$, indicating a positive correlation between the two variables, and the R-squared value was stronger than it was when analyzing MultiVAE ($R^2 = 0.164$). Further, a statistical significance could be shown when analyzing SLIM, as the p value was lower than 0.05 ($p = 0.012$). This showed that in the case of SLIM, familiarity does have a significant impact on perceived fairness.

Lastly, the impact of familiarity on perceived fairness was analyzed for ItemKNN. For this algorithm, the F-value was lower compared to SLIM, but higher in comparison to MultiVAE, $F(8, 106) = 1.533$. The same was seen for the R-squared value ($R^2 = 0.104$). No statistical significance could be shown in this case, as the p value exceeded 0.05 ($p = 0.154$). Thus, it was shown that familiarity does not have an impact on perceived fairness when it comes to ItemKNN.

4.2.3 The Effect of Satisfaction on Perceived Fairness

To answer RQ 3, whether satisfaction affects how users perceive fairness in the context of popularity bias, a one-way analysis of variance (ANOVA) was once again used. As with familiarity, the different algorithms were analyzed independently, and the threshold for statistical significance was set to $\alpha = 0.05$. In this case, the dependent variable was perceived fairness, while the independent variable was satisfaction.

The effect of satisfaction on perceived fairness was first analyzed for the MultiVAE algorithm. The analysis showed a positive F-value, $F(9, 105) = 5.694$, indicating a positive correlation between satisfaction and perceived fairness. As for the R-squared value, $R^2 = 0.328$. A statistical significance for the correlation could be shown, as the p value was less than 0.05 ($p < 0.001$). This indicated that satisfaction has a significant effect on perceived fairness for the MultiVAE algorithm.

Subsequently, the SLIM algorithm was analyzed. A positive correlation between satisfaction and perceived was shown, and the positive correlation was stronger than for MultiVAE, $F(10, 104) = 6.352$, $R^2 = 0.379$. Consequently, a statistically significant effect was shown in this case as well, as the p value was lower than 0.05 ($p < 0.001$). This showed that in terms of SLIM, satisfaction does have a significant effect on perceived fairness.

Finally, the effect of satisfaction on perceived fairness was analyzed for ItemKNN. In comparison to MultiVAE and SLIM, this algorithm had the highest F-value as well as the highest R-squared value, $F(10, 104) = 8.030$, $R^2 = 0.436$. This indicates that the positive correlation between satisfaction and perceived fairness was the strongest in this case. Consequently, a statistical significance could also be shown for ItemKNN ($p < 0.001$). Hence, it was shown that satisfaction had a significant effect on perceived fairness in this case as well. In summary, it was shown that satisfaction has a significant effect on perceived fairness for all of the tested algorithms.

5 Discussion

The purpose of this thesis was to understand the relationship between algorithmic and perceived fairness, through comparing users' perceptions of fairness to the algorithmic fairness of three different recommended playlists. Further, it aimed to gain knowledge on whether familiarity or satisfaction has an impact on perceived fairness. Based on this, three research questions were formulated:

RQ 1: Does perceived fairness correlate with algorithmic fairness in recommended playlists?

RQ 2: Does familiarity with the recommended music items in a playlist affect how users perceive fairness?

RQ 3: Does satisfaction with a recommended playlist affect how users perceive fairness?

To answer these questions, a survey was conducted. In the following section, the result of this survey is discussed in relation to the purpose and the research questions. In section 5.2, the method used in this study is discussed and evaluated.

5.1 Results Discussion

In previous sections, it was discussed how the topic of this thesis has not yet been explored, along with the need for it to be researched. This means that the results of this study cannot be related to much previous research, and in order to draw strong or decisive conclusions, more research is still needed. However, the results of this study can still be a starting point which can be built upon, challenged, and serve as a foundation for more in-depth research.

5.1.1 Algorithmic vs. Perceived Fairness

The first research question in this study aimed to understand whether a relationship between algorithmic and perceived fairness exists in the context of popularity bias. As described in section 4.2.1, the data showed that there was no significant correlation between perceived and algorithmic fairness. This means that even though the recommended playlists differed in their level of fairness, there was no clear difference in how the participants perceived the lists. Thus, participants with less fair playlists did not perceive them as being less fair than participants with fairer playlists and vice versa. Further, it was shown that even though the algorithms varied in their level of fairness, the participants could not tell any difference between them. No algorithm was perceived as being fairer than the others.

What this shows is that differences in fairness, in terms of including a fair balance of popular and less popular items in recommended playlists, are not noticed by users. This

could be a result of users not thinking much about the distribution of popular and less popular items, or even that they do not care much about it. Thus, it could be that this type of fairness is not important to them. This conclusion is in line with the study conducted by Sonboli et al. (2021), which showed that provider fairness is something that few people have ever thought about. If popularity bias has never been considered or thought about by users, it might be that it is an issue that is not of high importance to them. This finding is also in accordance with McNee et al. (2006), who state that users do not care about which algorithm performs the best based on different criteria – what the users want is good recommendations.

Based on this, it may be the case that the user perspective on popularity bias is not the most important aspect to consider when trying to mitigate it in recommender systems. Consequently, other perspectives might be of higher importance, such as the perspective of providers. If users of recommender systems cannot tell the difference between fair and unfair algorithms, in terms of recommending both popular and less popular items, it might be more important to focus on other factors, such as user satisfaction, as opposed to how they perceive this type of fairness.

5.1.2 Familiarity and Perceived Fairness

The purpose of the second research question in this study was to understand whether familiarity has an effect on perceived fairness in the context of popularity bias. The results showed that there was a positive correlation between the two variables for all of the algorithms. However, there was only a statistically significant effect when it came to the sparse linear method (SLIM).

In this case, SLIM was the only algorithm where familiarity had a significant effect on perceived fairness. When performing test runs of the study, it was noticed that SLIM tended to recommend the same artist multiple times. This resulted in playlists which were less diverse when it came to artists in comparison to the playlists generated by the other algorithms.

On the final page of the survey, participants were asked what influenced their opinion on whether a music item was popular or not – the song, the artist, or the genre. As described in section 4.1.2, most participants considered the song, followed by the artist. Genre was the factor that the lowest number of participants considered.

Based on this, we can assume that when the participants were not familiar with a song, they considered the artist to judge popularity. As SLIM tended to be less diverse in terms of recommending different artists, the perceived unfairness may have become stronger when the same artists were recommended multiple times.

What can be concluded is that the effect of familiarity depends on which algorithm the users are presented with. If the algorithm generates less diverse recommendations in

terms of artists, and the users are not familiar with the presented songs, it may be considered as being more unfair. As the other two algorithms seemed to be more diverse in their recommendations, perceived fairness was not affected by familiarity.

5.1.3 Satisfaction and Perceived Fairness

The third and final research question aimed to answer whether satisfaction has an effect on perceived fairness in the context of popularity bias. The analysis of the collected data showed that there exists a significant, positive correlation between satisfaction and perceived fairness. This means that an increase in satisfaction results in an increase in perceived fairness. The positive correlation was seen across all the three algorithms employed in the study.

What this shows is that the more satisfied users are with a recommended playlist, the fairer they consider it to be. This result is in accordance with the studies conducted by Woodruff et al. (2018) and Wang et al. (2020), who showed that satisfaction with an algorithmic outcome can have an impact on how we think about fairness. Thus, satisfaction biases our views on fairness.

This finding strengthens the suggestion that fairness, in terms of an algorithm generating a fair mix of popular and less popular items, is not of high importance to users, as satisfaction has a significant effect on how users think about it. Once again, it may be that other factors, such as satisfaction, are more important to users than this type of fairness.

5.2 Method Discussion

For this study, a survey was the method of choice. This method was found to be appropriate due to the need of quantitative data. The choice of conducting the study online enhanced reliability, as the delivery of the survey and questions was the same for all participants. Moreover, a sufficient number of participants were recruited.

The online platform and the questionnaire were reviewed and tested by researchers within the field of recommender systems to ensure validity. The survey was also tested by people with no prior knowledge of the study. This ensured that the platform was easy to use, and that the provided information as well as the questions were understandable.

No definitions of fairness or popularity bias were provided during the survey. Moreover, the included control questions helped identify fake and careless responses, which also enhanced the validity of the study. Moreover, the satisfaction questions in the questionnaire were adopted from Graus and Ferwerda (2021), who had already shown that the questions measured the same construct. This contributed to an increase in reliability of the questionnaire. Based on these factors, it can be argued that the

validity and reliability of the study are high. Thus, the results of the study are of value, and may be used as a foundation for further studies.

However, there are uncertainties regarding some aspects of the method. Firstly, it was shown that all of the fairness questions did not work very well, and in the end, only one fairness question was used to analyze the data. What the flaws in the questions were cannot be identified at this point, but it seems that qualitative research is needed to understand how the questions are interpreted by participants.

Further, it can be argued that the fairness questions may have measured satisfaction as opposed to perceived fairness. The term ‘fair’ was excluded from the questions to make them easier to understand. However, excluding this term might have led to the participants not thinking about fairness at all, but rather about if they were satisfied with the balance between popular and less popular items. Once again, qualitative research would have been needed, prior to conducting the survey, to understand which construct was being measured.

6 Conclusions and Further Research

In this chapter, the conclusions from this work are presented, as well as the practical and scientific implications. Lastly, suggestions for further research are proposed.

6.1 Conclusions

The purpose of this study was to examine the correlation between algorithmic and perceived fairness in the context of recommender systems. The study was conducted in a music recommender setting, and the fairness aspect investigated was popularity bias. In order to gain insights on this topic, users' perceptions of the fairness of different recommended playlists were compared to the algorithmic fairness of the playlists. Additionally, it was explored whether familiarity and satisfaction had a significant impact on perceived fairness. By conducting an online survey, the three research questions posed in this thesis could be answered.

After the collected data was analyzed, it was concluded that there is no correlation between algorithmic and perceived fairness. When analyzing the algorithms in isolation, it was shown that the participants could not tell a difference between playlists that varied in terms of fairness. Further, the algorithms used in the study differed in their level of fairness, in terms of displaying both popular and less popular items. Despite this, the participants could not notice a difference between them. These findings imply that fairness, in this context, may not be something that users find important.

Further, it was shown that familiarity only had a significant effect when it came to one out of the three algorithms, namely SLIM. What was noticed during test runs of the survey was that this algorithm tended to be less diverse in terms of recommending a variety of artist in comparison to the other algorithms. Hence, it recommended one artist multiple times in the same list.

The most common factor that participants relied on when judging popularity was the song, followed by the artist. It can be assumed that participants therefore looked at the artists when not knowing the songs, and as SLIM recommended many songs of the same artist, the perceived unfairness may have become stronger.

The other two algorithms were more diverse in their recommendations, and therefore, perceived fairness was not affected by familiarity. What can be concluded is that the effect of familiarity depends on other characteristics of the algorithm that the users are presented with. If the algorithm does not generate diverse recommendations in terms of artists, and the users are not familiar with the recommended songs, it may be considered as being more unfair.

When it comes to satisfaction, it was demonstrated that this is a factor that has a significant impact on perceived fairness. This was shown across all of the algorithms.

This finding indicates that satisfaction biases our views on fairness. Thus, it strengthens the argument that fairness may not be of high importance to users in this context.

6.1.1 Practical implications

Recommender systems are prevalent in many aspects of our daily lives, and enhancing their performance is of interest to society as a whole. The results of this study indicate that how users perceive fairness, in terms of a music recommender system generating playlists consisting of both popular and less popular items, might not be of high importance. Thus, when developing fair recommender systems, other stakeholders' opinions might be more valuable.

A stakeholder that has been shown to be negatively affected by popularity bias is the provider. Mitigating popularity bias according to providers' views is thereby of high importance, as the changes made can remain unnoticed by users. This could result in a positive effect for less popular providers, who could get their music recommended more often without negatively affecting the users. Thus, creating recommender systems that mitigate popularity bias can benefit the providers without negative effects for the other parties. This would create more opportunities for more providers, as their livelihood is in many ways dependent on exposure and being recommended (Patro et al., 2020).

Mitigating popularity bias and satisfying more providers would also be positive for the system, that is, the platform where the recommender system operates. Keeping the providers satisfied is crucial for making profit. The more providers that are satisfied with and choose to use the system, the better.

In terms of adapting recommender systems to the wants and needs of users, other factors may be more important to consider, such as satisfaction. As it was shown that satisfaction has an impact on perceived fairness, it can be argued that this factor is more significant when developing user-centric recommender systems. Moreover, diversity has been argued to affect the user experience in a positive way (Ferwerda et al., 2017; Kim et al., 2021). Diversity, in this case, could mean mitigating popularity bias and include more varied recommendations in terms of including both popular and less popular music items.

6.1.2 Scientific implications

When it comes to implications for the scientific community, it could be argued that it is important to consider different factors when researching and evaluating users' perceptions on fairness, such as satisfaction and familiarity. This might be the case in other contexts, and not only when it comes to popularity bias in music recommender systems.

The results of this study highlight the importance of taking other stakeholders into account when researching perceived fairness. As perceived fairness can be argued to

not be of high importance to users, other stakeholders' views should be considered next. Providers, who are negatively affected by popularity bias, could be one stakeholder that should play a bigger role in research regarding popularity bias.

As already mentioned, this study is a starting point in an area that remains relatively unexplored. Thus, there is room for more research to either confirm or reject the findings of this study. Fairness in recommender systems is an important field, as it has implications for all the stakeholders involved. Continuous research is therefore encouraged, as it is needed to help the industry develop fair recommender systems.

6.2 Further Research

For further research, the same study could be performed but on larger scale, including more participants, to either confirm or reject the results. New fairness questions can also be developed and used for a similar study, and it is suggested that these questions are researched qualitatively to ensure that they measure the intended construct. A similar study could also be performed using other algorithms, which perhaps differ even more in terms of algorithmic fairness. Additionally, it would be interesting to see if these results apply to other recommender domains, and not only when it comes to music recommender systems.

Another suggestion for further research is to conduct the same type of study but investigate perceived fairness from a provider perspective. Providers may have another perception of fairness, in terms of popularity bias, as they are affected by it in a different way than users are. Qualitative studies could also be performed, with interviews instead of surveys, to get more in-depth answers and understand why and how participants perceive fairness in a certain way. Moreover, the impact of other factors on perceived fairness could be researched.

Lastly, it is worth highlighting that popularity bias is only one out of many different types of (un)fairness. Although being a prevalent issue when it comes to recommender systems, other aspects are also of high importance. To conclude, perceived fairness needs to be investigated in various contexts, in terms of different aspects, and while considering multiple stakeholders.

7 References

- Abdollahpouri, H. (2019). *Popularity Bias in Ranking and Recommendation* Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA. <https://doi.org/10.1145/3306618.3314309>
- Abdollahpouri, H., & Burke, R. (2019). Multi-stakeholder Recommendation and its Connection to Multi-sided Fairness. *ArXiv, abs/1907.13158*.
- Abdollahpouri, H., Burke, R., & Mansoury, M. (2020). Unfair Exposure of Artists in Music Recommendation. *ArXiv, abs/2003.11634*.
- Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). *Recommender Systems as Multistakeholder Environments* Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, Bratislava, Slovakia. <https://doi.org/10.1145/3079628.3079657>
- Abdollahpouri, H., & Mansoury, M. (2020). Multi-sided Exposure Bias in Recommendation. *ArXiv, abs/2006.15772*.
- Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). The Unfairness of Popularity Bias in Recommendation. *ArXiv, abs/1907.13286*.
- Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020). The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. *Fourteenth ACM Conference on Recommender Systems*.
- Allen, M. (2017). *The SAGE encyclopedia of communication research methods*. SAGE publications.
- Andres, L. (2012). Validity, reliability, and trustworthiness. *Andres, L. Designing & doing survey research*, 115-128.
- Bauer, C. (2019). Allowing for equal opportunities for artists in music recommendation. *ArXiv, abs/1911.05395*.
- Bauer, C., Kholodylo, M., & Strauss, C. (2017). Music Recommender Systems: Challenges and Opportunities for Non-Superstar Artists. In *Proceedings of 30th Bled eConference* (pp. 21-32). <https://doi.org/10.18690/978-961-286-043-1.3>
- Boone, H. N., & Boone, D. A. (2012). Analyzing Likert Data. *The Journal of Extension*, 50.
- Burke, R. (2017). Multisided Fairness for Recommendation. *ArXiv, abs/1707.00093*.
- Celma, Ò., & Cano, P. (2008). *From hits to niches? or how popular artists can bias music recommendation and discovery* Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition, Las Vegas, Nevada. <https://doi.org/10.1145/1722149.1722154>
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020). Bias and Debias in Recommender System: A Survey and Future Directions. *ArXiv, abs/2010.03240*.
- Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, 8(1), 15951. <https://doi.org/10.1038/s41598-018-34203-2>

- Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogín, A., & Di Noia, T. (2021). A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction*, 31(3), 457-511. <https://doi.org/10.1007/s11257-020-09285-1>
- Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogín, A., & Noia, T. D. (2019). Recommender Systems Fairness Evaluation via Generalized Cross Entropy. *ArXiv*, *abs/1908.06708*.
- Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., & Pera, M. S. (2018). *All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness* Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v81/ekstrand18b.html>
- Elahi, M., Abdollahpouri, H., Mansoury, M., & Torkamaan, H. (2021). Beyond Algorithmic Fairness in Recommender Systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 41–46). Association for Computing Machinery. <https://doi.org/10.1145/3450614.3461685>
- Elahi, M., Jannach, D., Skjærven, L., Knudsen, E., Sjøvaag, H., Tolonen, K., Holmstad, Ø., Pipkin, I., Throndsen, E., Stenbom, A., Fiskerud, E., Oesch, A., Vredenberg, L., & Trattner, C. (2021). Towards responsible media recommendation. *AI and Ethics*. <https://doi.org/10.1007/s43681-021-00107-7>
- Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., & Getoor, L. (2018). A Fairness-aware Hybrid Recommender System. *ArXiv*, *abs/1809.09030*.
- Ferraro, A., Serra, X., & Bauer, C. (2021a). Break the Loop: Gender Imbalance in Music Recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (pp. 249–254). Association for Computing Machinery. <https://doi.org/10.1145/3406522.3446033>
- Ferraro, A., Serra, X., & Bauer, C. (2021b). What Is Fair? Exploring the Artists' Perspective on the Fairness of Music Streaming Platforms. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen, *Human-Computer Interaction – INTERACT 2021* Cham.
- Ferwerda, B., Graus, M. P., Vall, A., Tkalcic, M., & Schedl, M. (2017). *How item discovery enabled by diversity leads to increased recommendation list attractiveness* Proceedings of the Symposium on Applied Computing, Marrakech, Morocco. <https://doi.org/10.1145/3019612.3019899>
- Gipps, C. (2011). *Beyond Testing (Classic Edition) : Towards a Theory of Educational Assessment*. Taylor & Francis Group. <http://ebookcentral.proquest.com/lib/jonhh-ebooks/detail.action?docID=958110>
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504-528. [https://doi.org/https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/https://doi.org/10.1016/S0092-6566(03)00046-1)

- Graus, M. P., & Ferwerda, B. (2021). The Moderating Effect of Active Engagement on Appreciation of Popularity in Song Recommendations. In K. Toeppe, H. Yan, & S. K. W. Chu, *Diversity, Divergence, Dialogue* Cham.
- Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science and Technology*, 7, 396-403.
- Kim, J., Choi, I., & Li, Q. (2021). Customer Satisfaction of Recommender System: Examining Accuracy and Diversity in Several Types of Recommendation Approaches. *Sustainability*, 13(11), 6165.
- Koutsopoulos, I., & Halkidi, M. (2018, 12-15 Aug. 2018). Efficient and Fair Item Coverage in Recommender Systems. 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech),
- Kowald, D., Schedl, M., & Lex, E. (2020). The unfairness of popularity bias in music recommendation: a reproducibility study. *Advances in Information Retrieval*, 12036, 35.
- Krosnick, J. (2017). Questionnaire Design. In *The Palgrave Handbook of Survey Research* (2nd ed., pp. 439-455). https://doi.org/10.1007/978-3-319-54395-6_53
- Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In *Survey measurement and process quality* (pp. 141-164). <https://doi.org/https://doi.org/10.1002/9781118490013.ch6>
- Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems – A survey. *Knowledge-Based Systems*, 123, 154-162. <https://doi.org/https://doi.org/10.1016/j.knosys.2017.02.009>
- Lamb, G. D. (2003). Understanding "within" versus "between" ANOVA Designs: Benefits and Requirements of Repeated Measures.
- Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., & Schedl, M. (2021). Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected? In *Fifteenth ACM Conference on Recommender Systems* (pp. 601–606). Association for Computing Machinery. <https://doi.org/10.1145/3460231.3478843>
- Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018). *Variational Autoencoders for Collaborative Filtering* Proceedings of the 2018 World Wide Web Conference, Lyon, France. <https://doi.org/10.1145/3178876.3186150>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 140, 55-55.
- Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). *Feedback Loop and Bias Amplification in Recommender Systems* Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland. <https://doi.org/10.1145/3340531.3412152>
- McNee, S. M., Riedl, J., & Konstan, J. A. (2006). *Being accurate is not enough: how accuracy metrics have hurt recommender systems* CHI '06 Extended Abstracts

- on Human Factors in Computing Systems, Montréal, Québec, Canada.
<https://doi.org/10.1145/1125451.1125659>
- Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). *Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems* Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy.
<https://doi.org/10.1145/3269206.3272027>
- Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., & Schedl, M. (2021). Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5), 102666.
- Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & SOCIETY*, 35(4), 957-967. <https://doi.org/10.1007/s00146-020-00950-y>
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE*, 9(2), e89642.
<https://doi.org/10.1371/journal.pone.0089642>
- Ning, X., & Karypis, G. (2011, 11-14 Dec. 2011). SLIM: Sparse Linear Methods for Top-N Recommender Systems. 2011 IEEE 11th International Conference on Data Mining,
- Patro, G. K., Biswas, A., Ganguly, N., Gummadi, K. P., & Chakraborty, A. (2020). FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proceedings of The Web Conference 2020* (pp. 1194–1204). Association for Computing Machinery.
<https://doi.org/10.1145/3366423.3380196>
- Payne, S. L. B. (2014). *The Art of Asking Questions: Studies in Public Opinion*, 3. Princeton University Press. <https://doi.org/doi:10.1515/9781400858064>
- Peterson, R. A. (2000). *Constructing Effective Questionnaires*. SAGE Publications, Inc.
<https://doi.org/10.4135/9781483349022>
- Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.
- Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences*. Holt, Rinehart and Winston.
- Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). *Item-based collaborative filtering recommendation algorithms* Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong.
<https://doi.org/10.1145/371920.372071>
- Schedl, M., Brandl, S., Lesota, O., Parada-Cabaleiro, E., Penz, D., & Rekabsaz, N. (2022). *LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis* ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany. <https://doi.org/10.1145/3498366.3505791>

- Schelenz, L. (2021). Diversity-aware Recommendations for Social Justice? Exploring User Diversity and Fairness in Recommender Systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 404–410). Association for Computing Machinery. <https://doi.org/10.1145/3450614.3463293>
- Sharma, J. K. (2012). *Business statistics*. Pearson Education India.
- Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98, 277-284. <https://doi.org/https://doi.org/10.1016/j.chb.2019.04.019>
- Shrestha, Y. R., & Yang, Y. (2019). Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms*, 12(9), 199.
- Smith, J., Sonboli, N., Fiesler, C., & Burke, R. (2020). Exploring User Opinions of Fairness in Recommender Systems. *ArXiv*, *abs/2003.06461*.
- Sonboli, N., Smith, J. J., Berenfus, F. C., Burke, R., & Fiesler, C. (2021). Fairness and Transparency in Recommendation: The Users' Perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 274–279). Association for Computing Machinery. <https://doi.org/10.1145/3450613.3456835>
- Steck, H. (2011). *Item popularity and recommendation accuracy* Proceedings of the fifth ACM conference on Recommender systems, Chicago, Illinois, USA. <https://doi.org/10.1145/2043932.2043957>
- Stray, J., Vendrov, I., Nixon, J., Adler, S., & Hadfield-Menell, D. (2021). What are you optimizing for? Aligning Recommender Systems with Human Values. *ArXiv*, *abs/2107.10939*.
- Tanner, K. (2002). Chapter 5 - Survey research. In K. Williamson, A. Bow, F. Burstein, P. Darke, R. Harvey, G. Johanson, S. McKemmish, M. Oosthuizen, S. Saule, D. Schauder, G. Shanks, & K. Tanner (Eds.), *Research Methods for Students, Academics and Professionals (Second Edition)* (pp. 89-109). Chandos Publishing. <https://doi.org/https://doi.org/10.1016/B978-1-876938-42-0.50013-7>
- Wang, R., Harper, F. M., & Zhu, H. (2020). *Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA. <https://doi.org/10.1145/3313831.3376813>
- Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. Paper 656). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3174230>
- Yalcin, E., & Bilge, A. (2021). Investigating and counteracting popularity bias in group recommendations. *Information Processing & Management*, 58(5), 102608. <https://doi.org/https://doi.org/10.1016/j.ipm.2021.102608>

8 Appendices

Appendix A: Survey Platform



Appendix B: Questionnaire

Appendix C: Questionnaire Responses

Appendix D: Familiarity Responses

8.1 Appendix A: Survey Platform

Screenshots of the survey platform (desktop version).



Information about the study

In this study, we want to understand people's opinions on different playlists. The playlists will consist of songs that are recommended to you based on your taste in music. The study will take around **10 minutes** to complete.

The study starts with a few questions about yourself and a short personality test, followed by some questions about your relationship with music. After that, you will be presented with three different playlists, and you will be asked to answer some questions about them.

Requirements

- You must be at least 18 years of age.
- You need to have a [Last.fm](#) account.
- Your listening history on [Last.fm](#) must include at least 50 unique songs that can be identified by the recommender systems.

Privacy policy

For us to be able to get information about your listening history, you need to provide us with your username on Last.fm. Your username will not be used for any other purposes other than fetching which songs you have listened to. The answers you provide will not be traced back to you, and they will not be used for any other purposes than research, neither will they be shared with anyone that is not involved in the study.

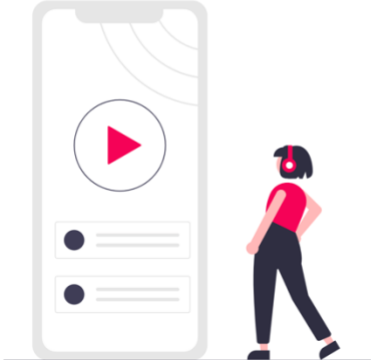
We urge you to be completely honest in your replies in order for us to get research data that is accurate and valid. Please note that participation is voluntary and that you can withdraw and exit the questionnaire at any time.

If you have read the above-mentioned information and want to take part in this study, please enter your username on Last.fm, and consent by clicking on the "Continue with the study" button.

Note: The survey includes mechanisms which detect fake contributions. Detection of such will result in exclusion and rejection of your MTurk Code.

Your username on Last.fm:

Continue with the study





Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se



10%

General information

Please fill in the following information about yourself.

Age:

Gender:

Country of origin:

Country of residence:

Continue

Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

47

Personality test

Below are a number of personality traits that may or may not apply to you. Please select an option to indicate the extent to which you agree or disagree with each statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

I see myself as:	Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
Extraverted, enthusiastic.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Critical, quarrelsome.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dependable, self-disciplined.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anxious, easily upset.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Open to new experiences, complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reserved, quiet.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sympathetic, warm.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Disorganized, careless.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Calm, emotionally stable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Conventional, uncreative.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Continue](#)

Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Your relationship with music

Below are some music-related statements. Please select an option for each statement.

Statement	Completely Disagree	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree	Completely Agree
I spend a lot of my free time doing music-related activities.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I enjoy writing about music, for example on blogs and forums.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I often read or search the internet for things related to music.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Answer with 'strongly agree' here.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I don't spend much of my disposable income on music.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Music is kind of an addiction for me - I couldn't live without it.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I keep track of new music that I come across (e.g. new artists or recordings).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Continue](#)

Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

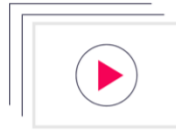
eveline.ingesson@ju.se

Song recommendations

On the following pages, you will be shown three different playlists with song recommendations that are personalized for you. For each playlist, you will be shown different statements and it is your task to select an option to indicate the extent to which you agree or disagree with each statement.

Next to each song in the playlist, there are checkboxes, where you will answer whether you know the song, have listened to the song, and if you think the song is popular.

Keep in mind when answering the questions that we are interested in your personal opinions, and that there are no right or wrong answers.

[Continue](#)

Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Playlist 1

Song title	Artist(s)	Genre	I know this song *		I've listened to this song		This song is popular	
thank u, next	Ariana Grande	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Carry Me Away	John Mayer	Rock	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
IDGAF	Dua Lipa	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Dance Monkey	Tones and I	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
break up with your girlfriend, i'm bored	Ariana Grande	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Eleven	Khalid	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Slow Dancing in a Burning Room	John Mayer	Rock	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Don't Start Now	Dua Lipa	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Gravity	John Mayer	Rock	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No
Eastside (with Halsey & Khalid)	Benny Blanco	Pop	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No	<input type="radio"/> Yes	<input type="radio"/> No

Questions

Please select an option for each statement.

Statement	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
I am satisfied with the list of recommended items.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would give the recommended items a high rating.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The list of recommendations matches my preferences.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Answer with 'disagree a little' here.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The list has a good balance between popular and less popular items.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The list would be more balanced if more popular items were included.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The list would be more balanced if more less popular items were included.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Is the ratio of popular items in this playlist the same as what you usually listen to?

- ☐ Yes, it is the same
- ☐ No, I usually listen to a lower ratio of popular items
- ☐ No, I usually listen to a higher ratio of popular items

[Continue](#)

Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Final Question

What influenced your decision on whether a music item was popular or less popular?
Check all the options that apply.

- ☐ The popularity of the artist
- ☐ The popularity of the song
- ☐ The popularity of the genre

[Complete survey](#)

Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Thank you for your participation!

Here is your unique code for Amazon MTurk: **CL2M8S1**



Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Thank you for your participation!

You have already participated in this study.



Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Unfortunately, you don't meet the requirements.

You don't meet the requirements to participate in this study.



Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

Unfortunately, you don't meet the requirements.

Your listening history on Last.fm doesn't meet the requirements for you to be able to participate in this study.



Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

You didn't pass the control question.

This survey includes control questions which detect careless contributions. Unfortunately, you didn't pass one of these control questions and can therefore not continue with the survey.



Eveline Ingesson

Department of Computer Science and Informatics

Jönköping University, Sweden

eveline.ingesson@ju.se

8.2 Appendix B: Questionnaire

Likert-type items used to measure satisfaction and perceived fairness.

Construct	Description	Source
Satisfaction	I am satisfied with the list of recommended items.	Graus and Ferwerda (2021)
Satisfaction	I would give the recommended items a high rating.	Graus and Ferwerda (2021)
Satisfaction	The list of recommendations matches my preferences.	Graus and Ferwerda (2021)
Perceived fairness	The list has a good balance between popular and less popular items.	–
Perceived fairness	The list would be more balanced if more popular items were included.	–
Perceived fairness	The list would be more balanced if more less popular items were included.	–

8.3 Appendix C: Questionnaire Responses

Tables showing the frequencies and percentages of responses for the Likert-type items used in the questionnaire. A score of 1 indicates the lowest possible level of perceived fairness, while a score of 5 indicates the highest possible level of perceived fairness.

Perceived Fairness of the Variational Autoencoder (MultiVAE)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	2.6	2.6	2.6
	2	10	8.7	8.7	11.3
	3	25	21.7	21.7	33.0
	4	48	41.7	41.7	74.8
	5	29	25.2	25.2	100.0
	Total	115	100.0	100.0	

Perceived Fairness of the Sparse Linear Method (SLIM)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	4	3.5	3.5	3.5
	2	9	7.8	7.8	11.3
	3	22	19.1	19.1	30.4
	4	52	45.2	45.2	75.7
	5	28	24.3	24.3	100.0
	Total	115	100.0	100.0	

Perceived Fairness of the k-Nearest Neighbors (ItemKNN)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	1.7	1.7	1.7
	2	8	7.0	7.0	8.7
	3	24	20.9	20.9	29.6
	4	51	44.3	44.3	73.9
	5	30	26.1	26.1	100.0
	Total	115	100.0	100.0	

Satisfaction with the Variational Autoencoder (MultiVAE)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	2	1.7	1.7	1.7
	2.33	2	1.7	1.7	3.5
	2.67	4	3.5	3.5	7.0
	3.00	5	4.3	4.3	11.3
	3.33	14	12.2	12.2	23.5
	3.67	24	20.9	20.9	44.3
	4.00	16	13.9	13.9	58.3
	4.33	30	26.1	26.1	84.3
	4.67	12	10.4	10.4	94.8
	5.00	6	5.2	5.2	100.0
	Total	115	100.0	100.0	

Satisfaction with the Sparse Linear Method (SLIM)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	.9	.9	.9
	1.67	1	.9	.9	1.7
	2.33	1	.9	.9	2.6
	2.67	6	5.2	5.2	7.8
	3.00	5	4.3	4.3	12.2
	3.33	14	12.2	12.2	24.3
	3.67	24	20.9	20.9	45.2
	4.00	14	12.2	12.2	57.4
	4.33	36	31.3	31.3	88.7
	4.67	8	7.0	7.0	95.7
	5.00	5	4.3	4.3	100.0
	Total	115	100.0	100.0	

Satisfaction with the k-Nearest Neighbors (ItemKNN)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	.9	.9	.9
	2.00	1	.9	.9	1.7
	2.33	2	1.7	1.7	3.5
	2.67	8	7.0	7.0	10.4
	3.00	4	3.5	3.5	13.9
	3.33	16	13.9	13.9	27.8
	3.67	13	11.3	11.3	39.1
	4.00	21	18.3	18.3	57.4
	4.33	35	30.4	30.4	87.8
	4.67	9	7.8	7.8	95.7
	5.00	5	4.3	4.3	100.0
	Total	115	100.0	100.0	

8.4 Appendix D: Familiarity Responses

Tables showing the frequencies and percentages of how many songs were known in each playlist generated by the different algorithms.

Familiar Songs in Playlists Generated by the Variational Autoencoder (MultiVAE)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4	3	2.6	2.6	2.6
	5	8	7.0	7.0	9.6
	6	14	12.2	12.2	21.7
	7	13	11.3	11.3	33.0
	8	16	13.9	13.9	47.0
	9	10	8.7	8.7	55.7
	10	51	44.3	44.3	100.0
	Total	115	100.0	100.0	

Familiar Songs in Playlists Generated by the Sparse Linear Method (SLIM)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	.9	.9	.9
	3	1	.9	.9	1.7
	4	3	2.6	2.6	4.3
	5	10	8.7	8.7	13.0
	6	13	11.3	11.3	24.3
	7	18	15.7	15.7	40.0
	8	12	10.4	10.4	50.4
	9	7	6.1	6.1	56.5
	10	50	43.5	43.5	100.0
	Total	115	100.0	100.0	

**Familiar Songs in Playlists Generated by k-Nearest Neighbors
(ItemKNN)**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	.9	.9	.9
	3	2	1.7	1.7	2.6
	4	2	1.7	1.7	4.3
	5	13	11.3	11.3	15.7
	6	15	13.0	13.0	28.7
	7	19	16.5	16.5	45.2
	8	10	8.7	8.7	53.9
	9	4	3.5	3.5	57.4
	10	49	42.6	42.6	100.0
	Total	115	100.0	100.0	