Understanding Context Effects in the Evaluation of Music Similarity

Michiel van den Berg

TU Delft

Understanding Context Effects in the Evaluation of Music Similarity

by

Michiel van den Berg

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday June 23, 2021 at 11:30 AM.

Student number:	4391039	
Project duration:	March 1, 2020 – June 23, 2021	
Thesis committee:	Prof. A. Hanjalic,	TU Delft, Responsible Full Professor
	Dr. J. Urbano,	TU Delft, Daily Supervisor
	Dr. S. Picek,	TU Delft

TU Delft

Preface

This thesis is the last part of my Master Data Science & Technology and the final step towards graduation. The process of the thesis was not as smoothly as planned, but throughout the process I learned a lot about both soft and hard skills. Doing a thesis in a global pandemic is unusual, but related to the process it has both benefits and disadvantages. I have not seen my supervisor and fellow students a lot in real life this year, which is quite unusual, but the benefit of meeting online was that code and documents can be shared on screen, which was in some situations very useful.

For the process of this thesis I want to thank a few people. First of all, I want to express my gratitude to my supervisor Julián Urbano for guiding me through this thesis. Especially the moments where I was lost in certain not relevant directions, you pushed me back on the main track. I learned thinking in a more systematic way, which I really appreciate. Also the discussions about hypothesizing context effects gave me new insights, which especially during this pandemic was very useful. I also want to thank my parents, brother and girlfriend. Although it sometimes took a bit long or they did not understand a thing about the explanation of the thesis, they kept supporting me till the end. The last people I want to thank are my closest friends. After a day of work, it was really nice to talk about the normal things in life and having my mind somewhere else than this thesis.

Michiel van den Berg
Delft, June 2021

Abstract

This work analyses context effect in the evaluation of music similarity performed by human annotators to better understand the impact of context effects in the current annotation protocol of Music Information Retrieval Evaluation eXchange (MIREX). Human annotators are known to be subjective when giving similarity judgements. The Audio Music Similarity task in MIREX uses human annotators to collect similarity judgements. The annotator gives judgements to a list of candidate songs that are similar according to the participating system. The annotation protocol has no clear guidelines, and on top of that, literature shows psychological effects which can influence the similarity score. Studies show that disagreement exists between different annotators in the Audio Music Similarity task. It is argued that the disagreement is due to the natural subjectivity of human annotators, but how much of the subjectivity is natural?

In this work, context effects are explored, which are the over- or underrating of candidate songs due to specific properties of the annotated list of candidates. The properties of the list of candidates are called factors and will be used as dependent variables. The exploration of context effects is split into two parts, 1) recognizing context effects and 2) measuring the impact of the context effect. New similarity judgements are collected through crowdsourcing, this data is checked on reliability before analysing the context effects. For recognizing context effects, the changes of previous judgements made by annotators are taken as a metric to see if the annotators are noticing potential context effects. The second part is measuring the magnitude of the over- or underrating by looking at the distance of the set of judgements to the ground truth. Hypotheses are made for the dependant variables change and distance, based on the factors Order, Trend, Location, Spread and Outlier. It seems that the collected data shows signs of context effects, with the Trend and Outlier hypotheses being in line with the data. The Order hypothesis seems to be the opposite of the data. When changes are made by an annotator, the final scores of the judgements are closer to the ground truth than before the changes. However, throughout the work, no significant results are found related to context effects.

Contents

List of Figures

List of Tables

1 Introduction

The time of buying music in a physical store is almost over. Nowadays music can be listened to on online platforms like Spotify and Youtube, with the benefit of widely accessible music. To make it easier for the music consumer to navigate through the large pool of online songs, Audio Music Similarity (AMS) can be used to automatically generate playlists, help a music consumer find new music, recommend songs, etc. To improve the systems using Audio Music Similarity, evaluation important and will be the focus in this work. In this chapter, the concept of Music Information Retrieval evaluation is explained, followed by an introduction to the evaluation protocols which involves human annotators. Then the points of criticism on the current evaluation methods involving human annotators are given. Finally, the motivation for this work is explained together with the approach to analyse them.

1.1. Evaluation in Music Information Retrieval

Evaluation is one of the key components in Information Retrieval research. Evaluation is the process of assessing how well a system meets the information needs of its users and therefore indicating how the retrieval process equates good performance with relevant system output [48]. The retrieval processes generally consist out of techniques, procedures, methods, or other processes which produce a list of items for a given statement or information need, also called a query. The goal of the retrieval systems is to produce a list of items which are the most relevant to the given query. To see if the list of items is relevant, evaluation is needed. The evaluation is not only indicating how the systems are performing, but also in which direction research should go. Throughout the years, evaluation is acknowledged as important in the development of information retrieval systems. The development of information retrieval systems usually follow a cycle which leads to better systems [47]. In this cycle, a task definition is defined to solve a research problem, with the task definition, a new system or adaption of a previous one is developed. In the evaluation phase, the system is assessed on how good it is. After evaluation, the results are interpreted and learned from, with the goal to either improve the system, or go back and improve the task definition by modifying it. Evaluation is a key component in this development cycle, leading to better quality systems.

Evaluation in Music Information Retrieval (MIR) is relatively scarce compared to the Text Information Retrieval but has become better with the introduction of the Music Information Retrieval Evaluation eXchange (MIREX) in 2005[1]. In 2005, 10 evaluation contests, also called tasks, were defined [10]. Throughout the years, this set of evaluation tasks is extended or modified, but in general, all tasks are following the so-called Cranfield paradigm [8]. This paradigm consists out of a document collection, a set of information needs, and a set of relevance judgements that describes the extent to which a document is relevant to an information need. The set of relevant documents is also called the ground truth. This ground truth is interesting for tasks where the ground truth is conducted by human an-

[1] https://www.music-ir.org/mirex/wiki

notators. The Audio Music Similarity (AMS) is one of these tasks and can be seen as one of the more general concepts of music similarity. A typical AMS system gives a list of songs to a given query based on the audio signal of the songs. The list of songs consists out of the most similar songs, according to the system, for a given query. To assess how similar the songs are to the query, human annotators are used.

The ground truth conducted by human annotators has a subjective nature, and it is this subjective nature which leads to some criticism on the way of collecting the ground truths for the tasks. The two points of criticism used in this work are 1) disagreement among different annotators, and 2) annotators going back to change previous judgements.

- **Disagreement:** The first point of criticism is based on the difference in score among different annotators, this phenomenon is also called the inter-rater agreement. Each annotator is giving a score based on what they find similar, which is subjective for each individual annotator. Therefore the annotators apply differently when making human similarity judgements [38]. In the 2006 MIREX edition, three annotators are used for each query-candidate pair, where the query is the input song for the system and the candidate one of the similar songs in the output list. Other editions of MIREX only used one annotator per query-candidate pair. It is indeed showed that different annotators give varying similarity scores [17, 25].

- **Change previous judgements:** The second point of criticism is the need for annotators to change their previous judgements [25], which has not been studied extensively in current literature but will be the main focus of this work. The reason for changing previous judgements during the process is the awareness of the annotator that previous judgements are over- or underestimated relative to the new judgements. The reason for this is probably due to the annotation protocol used in MIREX. This protocol consists out of a list of items which are annotated by the same annotator. The list of items is the whole list of songs which are similar to the query according to all participating systems. Annotating a list of items by the same annotator can lead to effects occurring due to the properties or conditions of the specific list of items being annotated. In this work, the occurring effects are called context effects.

The disagreement between annotators is not the main focus of this work, but will be discussed in chapter 4, to see if the collected data is reliable. The context effect will be the main focus and will be analysed in chapter 5.

1.2. Context Effects

The context effects in this work can be described as the over- or underrating of candidate songs due to specific properties or conditions of the list of candidates. It has been shown that disagreement exists among different annotators. When this disagreement is the result of a natural subjectivity, there is no need to eliminate or correct the judgements because they represent the real world. However, it is not exactly known how much of the disagreement is natural. It is claimed that there exists an upper bound for the performance of automatic analysis systems due to the disagreement among annotators. This upper bound has already been reached and not surpassed since 2009 for the AMS task[17], although not everyone agrees with this claim. The upper bound depends on many factors which could influence the decision making of the judges, think of past experience, musical background, personal preferences, and many more factors.

In the AMS task, the human annotator gets a list of candidates which need to be annotated within a bounded scale from 0 (not relevant) to 100 (relevant). Literature shows psychological effects when human annotators are giving a list of relevance judgements. Suppose an annotator is judging a list of candidates which order of appearance is low to high, then the annotator will probably give each candidate a higher score relative to the previous candidate. When the first candidate is already given a high score, the next candidates will be even higher because the candidate appearance is low to high. With the bounded scale of 100, the annotator could have a problem when annotating towards the edge of the scale. When this happens, the annotator could either change the previous candidates or correct

the next candidates within the boundaries of the scale. This is an example that will be seen as a context effect in this work. Not only the order, but also other properties of evaluated lists of candidates, such as Trend, Location, Spread and Outlier, will be taken as factors.

In this work, the changes made by annotators as well as the impact of the not corrected judgements will be studied. Changes have not been studied for the MIREX data, but it is shown that changes are made by annotators [25]. Why these changes are made is not explained. When context effects exist, it will be a sign that the differences between annotators are not a result of the natural subjectivity, but due to the current annotation protocol used in MIREX. This will lead to unreliable annotations and in the end, unreliable evaluations.

To eliminate context effects in the current annotation protocol, alternatives to the protocol have already been studied. Alternatives can be that not the whole list of candidates is annotated, but for example, preference judgements are made [46, 50]. Protocols where the annotator is judging a small set of candidates will probably eliminate the context effects.

1.3. Research Questions and Contribution

This work will be focusing on recognizing context effects by looking at the changes made by annotators and the differences in factors, as well as the impact of the context effects on the judgements. Therefore, the research questions are:

RQ1 *Are context effects measurable in the current annotation protocol of MIREX?*
To understand if context effects due to the current annotation protocol have an impact on the evaluation, they need to be recognizable in the data.

RQ2 *What is the impact of context effects in the current annotation protocol?*
Context effects can be seen as a problem when they have an impact on the annotations and in the end the evaluation of systems. When signs of context effects exist the magnitude of these effects have to be measured.

In order to investigate if context effects are recognizable and have an impact on judgements, new data is needed. A total of 120 querysets are created consisting out of the input of the systems, which is the query, and the output of the systems, which is a list of 15 candidates to be annotated. The set is as varying as possible with respect to the factors analysed.

The annotators used for collecting judgements are approached with the use of Amazon's Mechanical Turk (MTurk), which is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes [2]. A web-based system is created to collect the music similarity judgements. The system is similar to the system used in MIREX such that the annotator has the same environmental factors. The reliability of crowdsourced data compared to the MIREX experts will be analysed. A ground truth is established by taking the average over the other annotators. For a specific candidate judged by annotator i, the golden Fine score will be the average Fine scores given by the other annotators where $j \neq i$. In this way, the comparison to the ground truth will not include the Fine score of the annotator itself.

The context effects will be based on the changes made by annotators. The change is used as a dependent variable with the factors as independent variables. The impact will be measured by the distance of the Fine scores to the ground truth, and this distance will be used to indicate the over- or underrating.

1.4. Outline

The outline of the work will be as follows. Chapter 2 will state the current literature related to this work with a more extensive explanation of the definitions in the introduction. Chapter 3 will explain the methodology, including the design decisions and data collection. Chapter 4 will discuss the reliability of the crowdsourced music similarity judgements. Chapter 5 describes the existence of context effects

[2] `https://www.mturk.com/`

and their impact. In chapter 6, the limitations for this work will be discussed, and the last chapter will conclude the work.

2

Background

This chapter will cover the current literature on evaluation in MIR, and in particular on the AMS task. First, the general evaluation process will be described and how this is done in MIREX, followed by an explanation of the AMS task, together with the potential problems of human annotators performing the AMS evaluation. The potential problems will be translated into context effects with literature used as explanation. The last part of the chapter will be devoted to crowdsourcing, which will be used to collect the similarity judgements in this work.

2.1. Evaluation in Information Retrieval

As already mentioned in the introduction, evaluation is a key component in the development of information retrieval systems. In the evaluation of information retrieval, two broad ways of evaluation are known, system evaluation and user-based evaluation. User-based evaluation measures the user's satisfaction with the system, while system evaluation focuses on how well the system can rank documents [48]. When going back to the definition of information retrieval, it seems that user-based evaluation is the best way of assessing user satisfaction. However, for user-based evaluation, there is a need for a large, representative sample of users of the systems, equally well-trained users, equally well-developed systems etc [24]. This makes the user-based evaluation expensive and complex, which leads to IR researchers using system evaluation [48]. With system evaluation, there is need for (i) a document collection; (ii) a set of information needs, also called queries; and (iii) a set of relevance judgements, which is also called the ground truth or golden judgements. This kind of evaluation is following the Cranfield paradigm which was carried out by Cyril Cleverdon in the 1960's [7, 8] and is seen as the basis of information retrieval evaluation. Another successful project was the SMART system by Gerard Salton [36] which adopted the Cranfield paradigm and has the goal to investigate the effectiveness and efficiency of automatic methods of retrieval of text [6].

The last part of the Cranfield paradigm, the set of relevance judgements, has a central role in the evaluation of information retrieval systems. Where the document collection and set of information needs are used by the system to give relevant output, the set of relevance judgements is indicating how relevant the output is. The relevance judgements are used to make conclusions, when interpreted in the wrong way, it can lead to unnecessary borders or wrong direction in the research of music information retrieval.

2.2. Music Information Retrieval Evaluation eXchange

Most of the research in information retrieval evaluation is done for text-based systems. For this work, Music Information Retrieval is relevant. Music information retrieval is a younger discipline than text information retrieval, with the consequence that evaluation in this field is scarcer. With the start of

Figure 2.1: General overview of the components in the AMS task.

the Music Information Retrieval Evaluation eXchange (MIREX) in 2005[1], which is the equivalent of the Text REtrieval Conference (TREC)[2], the awareness of the importance of evaluation in MIR was improved. TREC and MIREX share many similarities and are both based on the standardization of the: 1) test collections of significant size; 2) tasks and/or queries to be performed on the test collections; and, 3) evaluation methods to be used to evaluate the results generated by the tasks/queries [12]. Despite the similarities, one major difference is the distribution of the datasets. MIREX is not able to distribute datasets to participants freely, due to the lack of freely available datasets in the current state of musical intellectual property copyright enforcement. This means that the participating systems are sent to MIREX and run by MIREX.

2.2.1. Audio Music Similarity and Retrieval

This work will focus on the Audio Music Similarity and Retrieval (AMS) task, which is one of the more general concepts of music similarity and is part of MIREX. The task is based on the Cranfield paradigm, where the components are:

- **Query:** The query is the information need and will be the input of the system. It is a randomly chosen song from each genre group.
- **Candidates:** The candidates are the output of the system, it is the list of songs that are the most similar to the query according to the system.
- **Similarity judgements** The list of candidates needs to be evaluated which is done by a human annotator. Each candidate is labeled with a score on how similar it is to the query.

An overview of the components can be found in figure 2.1. For this work, the last part is the most important. The set of similarity judgements is also called the ground truth. The ground truth for the AMS task is created by human similarity judgements which are collected with a web-based system called the Evalutron 6000 (E6K) [19], developed for MIREX tasks. The E6K collects similarity judgments for query-candidate pairs generated by the submitted algorithms. For each query-candidate pair, judges are asked to input two similarity evaluations: 1) a Broad category of similarity (i.e., Not Similar (NS), Somewhat Similar (SS), and Very Similar (VS)); and, 2) a Fine score between 0.0 (Least similar) and 10.0 (Most similar). The Fine score could also be between 0.0 (Least similar) and 100.0 (Most similar), depending on which year. The Fine score has an expected higher variability and is more likely to be prone to context effects. Due to this expected higher variability, only the Fine scores are used in this work.

At the start of MIREX in 2005, the AMS task was not part of the evaluation exchange. The AMS task started in 2006 but has not been part of the exchange every year. The AMS task was part of MIREX in the years from 2006, 2007, 2009 till 2014. Not each year has the exact same protocol concerning the amount of queries, candidates and annotators, for example, 2006 was the only year with three different annotators per query-candidate pair. Also the amount of queries changed in 2012 from 100 to 50, however, in the same year the amount of candidates changed from 5 to 10. The overall amount of judgements to be made per system is therefore the same, except for the 2006 edition. See table 2.1 for an overview, where the values are based on one participating system.

[1] `https://www.music-ir.org/mirex/wiki`
[2] `https://trec.nist.gov/overview.html`

Year	Queries	Candidates	Annotators
2006	60	5	3
2007	100	5	1
2009	100	5	1
2010	100	5	1
2011	100	5	1
2012	50	10	1
2013	50	10	1
2014	50	10	1

Table 2.1: The amount of queries, candidates per query and annotators for the AMS task in each year. The table is based on one participating system.

The queries are randomly chosen from a collection of songs in certain genres. The genre can be important for similarity judgements, because if the candidate is from the same genre as the query, it is easier to give a higher score than another genre. But also the variety in music songs is important to include in the evaluation of the systems. The systems have to be evaluated for all music, not only for a particular genre. In 2006, the queries were skewed towards Rock/Pop, roughly 50% of examples are labeled as Rock/Pop, while a further 25% are Rap & Hip-Hop[3]. After the 2006 edition, the queries were equally representing ten different genre groups: Baroque, Country, Edance, Jazz, Metal, Raphiphop, Rockroll, Romantic, Blues, Classical.

2.3. Human Annotators

Information retrieval can be described as finding material of an unstructured nature that satisfies an information need from within large collections [30]. This definition can be interpreted as the process of searching through a set of documents to find items that may help to satisfy the information need. However, 'satisfies the information need' can be seen a something which is subjective, especially when humans are involved in the process. Human behaviour and human thinking is complex, which makes it difficult for systems to satisfy the information need for each individual user at each point in time. When items are retrieved from a set of documents, not all users will agree on how relevant these items are. In other words, there is some subjectivity involved.

2.3.1. Challanges of Music Similarity

When considering music, this subjectivity is even more complex because music is multidimensional. There are a lot of factors which can influence the way music is interpreted, but also the way music is represented can cause challenges in the development and evaluation of MIR systems. J. Stephen Downie [11] describes several challenges for the field of music information retrieval. First of all, there are multiple facets in music, think of pitch, temporal, harmonic, timbral, editorial, textual, and bibliographic facets. When considering the similarity of music, these different facets can be taken into account individually. But in practice, these facets are not mutually exclusive, meaning that the interaction of different factors also plays a role in similarity. Downie describes this challenge as the multifaceted challenge. Another challenge is the way music is represented, which could be symbolic or melodic. And even in these two categories, multiple ways are possible. For example, the symbolic representation of music can be notes, text, encoding etc. Melodic could be live performance, acoustic, analog, digital etc. Downie describes this challenge as the multirepresentational challenge. Music is also perceived differently across people. The experience of music will vary among people based on their mood, situation, and circumstances, also called the multiexperiential challenge. Other challenges Downie describes are the multicultural challenge and the multidisciplinarity challenge. The

[3] https://www.music-ir.org/mirex/wiki/2006:Audio_Music_Similarity_and_Retrieval_Results

described challenges can be summarized as subjectivity in music similarity. When looking at evaluation in MIR, it is exactly this subjectivity which makes it difficult to evaluate information retrieval systems. But at the same time, evaluation of these systems can be seen as a key component in the development of information retrieval systems.

2.3.2. Disagreement among Annotators

For the AMS task in MIREX, the similarity judgements are given by human annotators which have some affiliation with MIREX. They can be seen as experts or at least be seen as annotators which have some knowledge about music similarity. However, it seems that between different "expert" annotators, there still is some disagreement about how similar a candidate is. In the 2006 MIREX edition, three annotators are used for each query-candidate pair, while the other editions only used one annotator per query-candidate pair. For the 2006 edition, it is indeed shown that different annotators are varying in Broad and Fine score, this difference is also called the inter-rater agreement. The inter-rater agreement for the Broad score in 2006 is calculated with Fleiss's Kappa [16], and gives a score of 0.2141 for the AMS task [25]. Fleiss's Kappa scores can range from 0.0 (no agreement) to 1.0 (perfect agreement), the AMS score above is considered as a fair level (0.21-0.40) of agreement [27]. But the score is at the low end of the fair level range which indeed confirms the subjectivity in human relevance judgements. The Fine score inter-rater agreement can not be calculated with the categorical Fleiss's Kappa, but Pearson correlation shows that the difference between the three different judges in 2006 is ranging from 0.37 to 0.43, which again is quite low [17]. For example, query-candidate pairs which are rated as very similar, a score between 9 and 10, by one grader in 2006, have on average a score around 6.5 by the other judges. The relatively low inter-rater agreement in the 2006 edition is confirming the subjective nature of human similarity judgements. A consequence of the low inter-rater agreement between annotators is the low chance of a perfect agreement, which is reached when all the judges give the same score to a system. This perfect agreement will probably never be reached due to the subjective nature of annotators. It is claimed that there exists an upper bound for the performance of automatic analysis systems which has already been reached and not surpassed since 2009 for the AMS task[17], however, not everyone agrees with this claim. The upper bound depends on many factors which could influence the decision-making of the annotators.

2.3.3. Changing Annotations

The decision-making is not only based on the factors related to a specific annotator, for example past experience, musical background, and personal preferences. There are also factors related to the annotation protocol used in MIREX. The annotators are given the whole list of candidates which needs to be annotated by the same annotators. They are able to take a break and come back to finish the list of candidates, but the principle of annotating multiple candidates within a short time period stays the same. In psychology, effects are described when annotators are judging a list of items. The magnitude of these effects is dependent on the characteristics of the queryset. For example, annotators are careful with their first judgements and will not give scores close to the ends of the scale, in order to leave some space for judging the next candidates [15]. When having a queryset of very similar candidates to the query, this means that the judgements will be underestimated due to staying away from the ends of the scale. Meaning that the judgements are underestimated or annotators are going back to change previous judgements. Another example would be the behavior of annotators when candidates are close to each other. Candidates are given a score based on the previous candidates [45], when close to each other, there is no space left to correct previous candidate scores, resulting in annotators going back and change these candidate scores. The AMS task in 2006 showed that people go back and change previous judgements, but the reason why is not clear [25]. In this work, the changes made by annotators are explored based on the properties of the queryset, also called the context effects.

2.4. Context Effects

Context effects are described as the influence of environmental factors on someone's decision-making. The environmental factors studied in this work are based on the properties of the querysets, the Fine

scores in the querysets together can indicate these properties. For example, the median of the set of Fine scores indicates where the set of scores is on the scale. The factors used are Order, Trend, Location, Spread, and Outlier and will be further explained in section 3.2. Context effects have not been widely studied, to my knowledge, in the field of music information retrieval. Therefore, this section will mainly be based on the text information retrieval field.

One of the best studied context effect is the order effect, which will be one of the factors in this thesis. The order effect can be described as an effect that occurs when results are not consistent if different orders are used. In other words, the presented order matters for the results. In the case of music similarity judgements, an order effect exists when it matters in which order the list of candidates is presented to the judge. Eisenberg and Barry [15] conducted a study related to the order effect back in 1988. In their experiment, they ask annotators to judge a list of relevance judgements, which is either in a high to low or low to high order. These orders are based on already been made relevance judgements which were conducted in random order by an earlier study of their own. They show that an order effect exists when different orders are used. Annotators tended to underestimate the relevance of documents when the order was presented in a high to low order, and the relevance of documents was overestimated when a low to high order was used. The study by Eisenberg and Barry [15] shows that an order effect exists, but the reasons why is not discussed in depth.

Besides the Order effect as main effect, the other factors can also influence the results and are chosen based on context effects found in literature. Not only the main effect of each factor but also the interaction effect of factors is interesting for the magnitude of the effect. The following context effects found in literature are used to define the factors: End-aversion, Anchoring effect, Cursoriness effect, Learning effect, Decoy effect, and the Diagnosticity effect.

- **End-aversion:** In one of the first studies by Eisenberg and Barry [15], they suggest that people are hedging their bets. They find it conceivable that subjects were reluctant to assign extremely high or low scores to the documents presented first. If a judge assigned a 1 to the first document and then found an even less relevant document, the scale does not provide a means of indicating that judgment. By assigning a 2 or 3 to the first document, the judge has maintained the option of judging following documents as either more or less relevant than the first document. This phenomenon is also referred to as end-aversion, meaning that people tend to stay away from the extremes. In the case of the Fine scale, it would mean that annotators will less likely choose the ends of the Fine scale at the beginning of the judge session, i.e. the first judgements will in most cases not be at the ends of the scale. And even when an annotator has judgements at the ends of the scale, other judges judging the same candidate are likely to have other ratings, causing the mean to be away from the extremes. Flexer [17] showed for the 2006 data, which have three judges per query-candidate, that when a judge has a judgement between 0-1, the other judges have an average of around 3. For the other end, when a judgement is between 9-10, the other judges will have an average of 6,5.

- **Anchoring effect:** The anchoring effect is another effect which can occur when an annotator is giving judgements. Anchoring describes the phenomenon that a given stimulus affects later judgments in the direction of the previous judgment, even if both stimuli are completely unrelated. Therefore people adjust their estimation towards an initial presented value. [45]. Also Strack and Mussweiller [42] explained that anchor values serve as the reference point for people to adjust their values which have to be estimated. One example of their study is asking people whether the age of Gandhi at his death was higher or lower than 9 years, or whether the age of Gandhi at his death was higher or lower than 141 years, the average guess of his actual age at death was lower when 9 years was used in the question. People, therefore, estimate their value towards an initial presented value. Both studies above are based on anchor values which are more extreme than the range of plausible answers. In the case of relevance judgements, the anchor value or values will not be extreme, because there is a limited Fine scale, and the anchor values are self-generated within the boundaries of that Fine scale. In a later study of Strack and Mussweiler [33], they study the anchoring effect with anchor values which are in the range of

plausible answers. They showed that judges search for ways in which their answer is similar to the anchor value, and thus are based on the first estimation value. This study will better fit the case of music similarity judgements which have to be rated within the Fine scale.

- **Cursoriness effect:** The cursoriness effect is also known as fatigue, which is based on energy and cognitive capacity consumption of judges who read or listen carefully to candidates, leading to a lower cognitive capacity for future readings. Without a break to replenish the cognitive capacity, fatigue sets in [51]. This means that judges pay less attention and motivation to the candidates at the end of the list of candidates compared to the beginning of the list. This phenomenon is also confirmed by Huang and Wang [22], where document sets of 75 documents showed fatigue among judges. The cursoriness effect is especially interesting for the order. Huang and Wang studied the relationship between the number of documents judged and the order. They use different list sizes of the to be judged documents to see for which number of documents an effect occurs. Significant effects exist when 15 and 30 documents are presented, 45 and 60 documents show signs of an effect but are not significant. When the number of documents was 5, no effect was found. Therefore, they conclude that list with fewer documents than 15 will not be influenced by an effect, which was already confirmed by a study of Purgailis Parker and Johnson [35] where they indicate that users are not influenced by an order effect when the set of retrieved documents is less than 15. Both studies assume that fatigue can be an important reason for large sets, but no clear indications are given for this assumption.

- **Learning effect:** The learning effect is kind of the opposite of the cursoriness effect. Xu and Wang [51] describe the learning effect as participants gaining more knowledge about documents over time. Therefore, documents presented later might be regarded with a lower degree of relevance, because the information need it fulfills has already been satisfied. This satisfaction is probably more related to relevance search, but the concept of the learning effect could also be applied to music similarity judgements. The concept of the learning effect used by Xi and Wang [51] is based on Sperber and Wilson's [41] relevance theory, according to which, "documents are relevant to a user because the new information is able to work together with the existing knowledge (i.e., cognitive context) to produce new knowledge, or to strengthen or weaken the confidence in the existing knowledge." In the case of music similarity judgements, it would mean that judges are learning throughout the session of judging all candidates of a query set. With the consequence of having biased judgement at the end of the session due to the increasing knowledge gain in the session.

- **Decoy effect:** The decoy effect is defined as a situation in which the addition of a new option to a choice set will make the target option in that set more attractive to consumers and more likely to be chosen. It is more relevant to choose one item out of a list of items, but can be relevant to a list of query-candidate pairs, where the most similar candidate has to be chosen. Euckhoff [13] showed that in crowdsourcing scenarios there is a considerable risk of suffering from Decoy Effects when multiple options are shown for relative ranking.

- **Diagnosticity effect:** Tversky [44] started with studying the effect of what he called diagnosticity, which is an effect when categorization has an influence on the item similarity. The idea is that features used as the basis for categorization acquire diagnostic value and increase the similarity of the objects that share them [43]. Considering the music information retrieval, it would mean that the similarity of a candidate is dependant on the other candidates in the set. This same candidate could have another level of similarity when judged in another set.

There is a lot more to find and study in the field of psychological effects on the decision-making process of people. The background is needed to create factors for a varying dataset to increase the chance of observing a context effect. However, the focus will not be on how the measured effects can be explained according to psychological literature, but on how and if the context effects are measurable in the data.

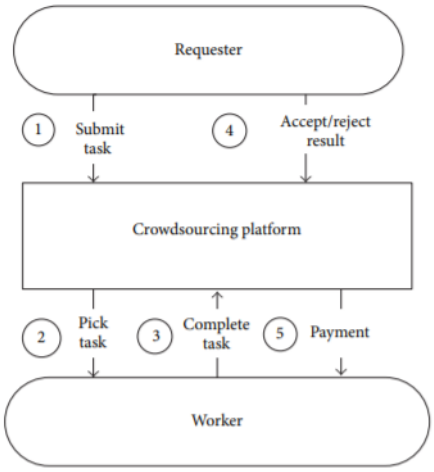


Figure 2.2: Flow of submitting and completing tasks via crowdsourcing.

2.5. Crowdsourcing

The similarity judgements collected in the MIREX AMS task are given by annotators which have a stake in Music Information Retrieval and Music Digital Library research. These people have a connection to the field and can be considered as experts when giving similarity judgements. For this work, not the experts, but the crowd is used to obtain similarity judgements. The phenomenon of using the crowd for problem solving is also called crowdsourcing. Howe [21] described crowdsourcing as the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. In the case of music similarity judgements, the judgements are not made by people which have a stake in Music Information Retrieval, but made by an undefined network of people.

In this work, crowdsourcing will be done using Amazon's Mechanical Turk (MTurk), which is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually[4]. With MTurk, a Human Intelligence Task (HIT) is created by the requester and this task is listed on the platform. Workers who meet the requester requirements can perform the task and based on the performed task, the requester can approve or reject the worker. Rejection will happen when the performed task does not pass the quality control measures of the requester, but this process should be done fairly. When the work is approved, the worker will be paid with the before mentioned compensation. See figure 2.2[37] for a flowchart representation of the crowdsourcing process.

2.5.1. Benefits and Risks

Crowdsourcing has the main advantage of having a relatively low cost compared to experts. Instead of experts judging the query-candidate pairs, individuals who have access to a device and have an internet connection can perform the task. The individuals are mostly unemployed and have a low level of education, but due to the open access, it is easy to perform the tasks. In the early years of crowdsourcing, the compensation of the tasks was not high which made it very cheap for the requester. Nowadays, the compensation is more ethical and around minimum wage, which is still relatively cost-efficient. Not only the low cost is beneficial, but crowdsourcing can be done in parallel with a lot of people, which makes it time-efficient. The MIREX AMS task takes around two weeks, table 2.2, while the same amount of tasks done with crowdsourcing is a matter of days or even hours [28, 29].

Crowdsourcing also has the benefit of having a diverse group of people [4, 37, 49]. People in the crowd can be diverse in language, culture, background, age, country etc. Related to the AMS task, this

[4] `https://www.mturk.com/`

	AMS
MIREX 2006	15 days
MIREX 2007	8 days
MIREX 2009	14 days

Table 2.2: Number of days needed to collect all similarity judgements

is beneficial because the experts are not the end-users of the system. When using a small group of annotators, the judgements can be tailored to this small group. When having a big diverse group of people, it better reflects a real-life scenario in which the systems will be deployed.

If there were only benefits with crowdsourcing, all research would be done using the crowd. However, this is not the case, there are also downsides when conducting experiments using the crowd. One of the major disadvantages of using crowdsourcing is the reliability of the experiments or data collected. In most cases, it is not known who the people in the crowd are and if they can be trusted. Although some platforms do have the option to select some of the prerequisites of workers using the platform, these are still general options. And even when the crowdworker has the prerequisites, it does not automatically mean that the worker will put effort into the task. The consequence of not completely trusting the crowd is the chance of not having reliable data. Having reliable data is a key component of making conclusions or directions in research.

2.5.2. Quality Control

To eliminate the chance of having unreliable data, quality control measures have to be taken to preserve the quality of the work. Almost all studies using crowdsourcing make use of quality control in some way. The most common ones found are:

- **Verifiable question:** A measure which is commonly used, is the use of a verifiable question before the task to see if the worker understands the question, and if the worker has the prerequisites to perform the task. The verifiable question can be a small task which is part of the real task to perform, but it could also be a question to check if the user is not a machine. All kinds of questions have the same purpose, which is beforehand filtering the users who will perform an unreliable task. [3, 26, 32, 37]

- **Repeating tasks:** When repeating the tasks, the same task will be performed by different annotators. When using multiple annotators for the same task, the task is not dependant or tailored to only one annotator. When a high agreement between different annotators occurs, it can be seen as reliable data. The other way around is the same principle, different annotators with no agreement, could indicate that the data is unreliable. The level of agreement is really dependant on which task or purpose the crowd is used, but in general, a higher level of agreement indicates better reliability. [2, 3, 20, 23, 28, 32, 34]

- **Response time:** The response time on the task by the worker is an easy to use measure. The only implementation needed for this measure is two timestamps or more when measuring time spent on sub parts of the task. When workers are not putting effort into the task, they will not spend a lot of time doing the task. This could be an indication that the results of the worker are unreliable. [26]

- **Worker reputation:** Most of the workers have done multiple tasks before which are approved or rejected. The benefit of using a crowdsourcing platform like MTurk, is the monitoring of the workers by MTurk itself. This gives the opportunity to see or choose which workers have a good reputation on the platform, i.e. which workers have a good approved-reject ratio. MTurk gives, for an additional fee, the option to choose workers who have the Master qualification. These workers have shown that they can provide reliable data. [14, 32, 37]

- **Ground truth comparison:** This measure cannot be applied in all situations, but when a ground truth is available, it can be used to compare the crowdsourced data. This is not a measure to filter unreliable workers beforehand, but a measure which is used after the data is collected. The data can be compared to the ground truth to see if the way of collecting data is leading to reliable data. Having reliable data is an important aspect for further analysis. [32, 34]
- **Trap questions:** The last common measure to discuss, is the trap question. This is a question which will check if the worker is paying attention or putting effort into the task. There are multiple ways to make use of this measure, related to similarity judgements, the requester could list a candidate twice to see if both scores are similar. Another way is to list the query in the list of candidates, this means that the query will be compared with the query itself leading to a maximal score. There are other forms of trap questions, but the main purpose is to filter out workers who are not paying attention. [28, 29, 37, 52]

3

Methodology

This chapter will explain the process of creating, collecting, and analysing music similarity judgements. The approach in this work is mainly quantitative with some qualitative methods for recognizing context effects. The creation of the dataset is based on the available MIREX data from previous years and based on factor levels which will be further explained in this chapter. The factor levels are used to get a varied dataset, which is important for measuring context effects. After explaining the creation of the instances, the procedure of collecting the similarity judgements will be explained, together with describing the collected data.

3.1. Dependent Variables

The measurement of the context effects will be split into two parts, 1) recognizing context effects and 2) measuring the impact of the context effect. To answer the first part, a metric is needed to see which querysets are prone to context effect. The metric will be the amount of changes made by the annotator. In a judgement session of an annotator, some will notice an effect, some will not. When an effect is noticed by the annotator, they will change their previous judgements and therefore adjust their previous over- or underestimation of the judgements. When not being noticed, this over- or underestimation is not corrected and therefore has some distance to the true value of the judgement. This distance is used to answer the second part, the impact of the context effects.

3.1.1. Calculation

To capture all kinds of change, six response variables are created to capture the different ways of looking at change. The response variables are Count, Total change, Average Total change, Direction of change, Average Direction of change, and Where the change is made. They can be described as:

- **Count:** The total amount of changes made in the queryset.
- **Total:** The absolute value of all changes in a queryset together.
- **Average Total:** The absolute value of all changes in a queryset together divided by the amount of changes made in the queryset. Total change divided by Count.
- **Direction:** The total value of positive changes substracted by the total value of negative changes in a queryset.
- **Average Direction:** The total value of positive changes substracted by the total value of negative changes in a queryset divided by the amount of changes made in thequeryset. Direction of change divided by Count.

- **Where:** Each candidate number is given a rank between 0-1 based order of appearance. The average rank of all changed judgements in a queryset is taken to indicate where the changes are made.

The distance is used to indicate the impact of the context effects. When annotators over- or underestimated their judgements, it means that the judgements have some distance to the ground truth. To indicate the distance to the ground truth, two metrics are used:

- **Pearson correlation:** The Pearson correlation is used to evaluate the linear relationship between two Fine scores and therefore taking proportions into account. It is calculated with the following formula:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{3.1}$$

  where X and Y are the sets of Fine scores of two annotators, cov is the covariance and $\sigma_x \sigma_y$ are the standard deviation of X and Y.

- **Root-Mean-Square Error** The Root-Mean-Square Error (RMSE) is used to indicate, for each Finescore in the set, the distance to the ground truth. It is calculated with:

$$RMSE = \sqrt{(\frac{1}{n}) \sum_{i=1}^{n} (y_i - x_i)^2} \tag{3.2}$$

3.2. Independent Variables

In order to measure context effects, factors needs to be considered which are interesting based on existing literature. These factors are used as the independent variables when analysing the data. The factors used are the Order, Trend, Location, Spread, and Outlier and can be described as:

- **Order:** The list of candidates in each queryset appears in a certain order to the annotators. The order of appearance is most interesting related to the anchoring effect, cursoriness effect, and learning effect. The Order factor will have three levels: High to Low (H2L), random and Low to High (L2H).

- **Trend:** The Order factor is evaluating the monotonic increase or decrease of the queryset. When having the list of Finescores itself, the trend of the Finescores can be calculated. Same as Order, the appearance order of the list of candidates is used, but now together with the Fine scores instead of the ranks. The Trend is different than the Order factor, but considering the psychological effects, they share similarities. The anchoring effect, cursoriness effect, and the learning effect are interesting. The reason to include Trend separately from the Order factor is the magnitude in which they can differ as well as the interaction effects with other factors. The levels of the Trend factor are: Exponential (Exp), Linear, Flat, Logarithmic (Log).

- **Location:** Another factor to consider is the Location. The Location factor is especially based on the end-aversion in literature. People tend to stay away from the extremes, which can be interesting for the location of the queryset. The location indicates how far away the queryset is from the extreme. The scale used for the Finescores is bounded between 0 and 100. Within the boundaries of the scale, the annotator is free to give judgements wherever they want. The Location will be using three levels: High, Middle and Low.

- **Spread:** The Spread factor is indicating the range of the Finescores in the queryset. The range is bounded between 0 and 100, but annotators can differ in how much of that scale they use. Considering the effects in literature, Spread is based on the decoy effect and the diagnosticity effect. The spread indicates what the range of the queryset is, i.e. how close are the candidates to each other. Candidates in a low spread are expected to be similar, while a high spread can indicate a varying set of candidates. The levels of the Spread factors are: High, Middle and Low.

- **Outlier:** The Outlier factor indicates whether or not the queryset contains outliers. A set of Finescores contains an outlier when mostly one Finescore is far from the other Finescores. The Outlier is chosen based on the decoy effect, diagnosticity effect, and the end-aversion. The outlier has some overlap with other factors because having an outlier means the rest of judgements are close together at the opposite level of the outlier. The levels of the Outlier Factor are: High, None and Low.

To illustrate when a list of candidates belongs to a certain factors, examples are shown in figure 3.1.

3.2.1. Calculation

The calculation and labeling of the factors is based on a queryset of 15 candidates, an overview of the calculation can be found in table 3.1. The calculation of the factors is done as follows:

- **Order:** The Order factor will have three levels: High to Low (H2L), random and Low to High (L2H). For calculating the Order, Spearman's rank correlation coefficient is used on the list of candidates, which is calculated with the following formula:

$$\rho = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{3.3}$$

  where rg_X is the list of Finescores of the candidates and rg_Y is the appearance order, both converted to ranks, cov is the covariance and $\sigma_{rg_X}\sigma_{rg_Y}$ are the standard deviation of rg_X and rg_Y. The correlation coefficient is used to label the queryset to a Order level with the following thresholds: H2L == $\rho < -0.2$, Random == $-0.2 < \rho < 0.2$ and L2H == $0.2 < \rho$. The reasons for using ranks instead of the Finescores itself is that the Order should be evaluated in a monotonic way. The trend of the Finescores itself will be used in the Trend factor.

- **Trend:** The levels of the Trend factor are: Exponential (Exp), Linear, Flat, Logarithmic (Log). For Labeling the queryset with a level, three formulas are fitted to the set of Finescores:

$$y = b_0 + b_1 * x \tag{3.4a}$$

$$y = x^b \tag{3.4b}$$

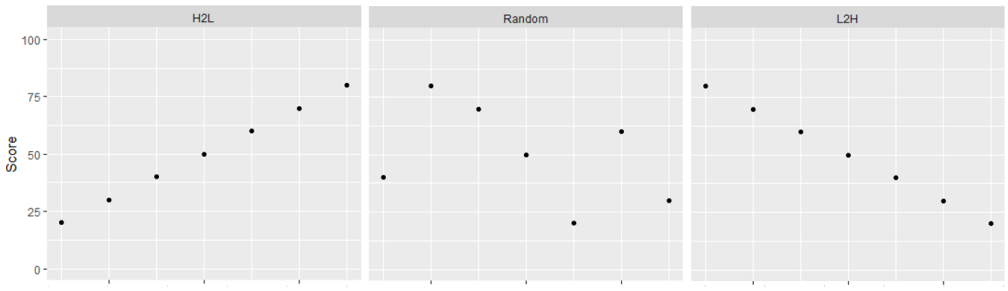
$$y = 1 - x^b \tag{3.4c}$$

  Where y is the set of Finescores and x the appearance order. The first formula is the linear model, the second one the exponential model and the third one is the logarithmic model. Each set of Finescores is fitted to all formulas and based on the residual sums of squares, the model which fits the set of Finescores the best is chosen. When the linear model is chosen, the queryset will be labeled as flat when $-2.2 < b_1 < 2.2$. This threshold is used based on the Finescore scale from 0 to 100. With 15 candidates the threshold indicates a flat set when the fitted model only uses one-third of the scale.

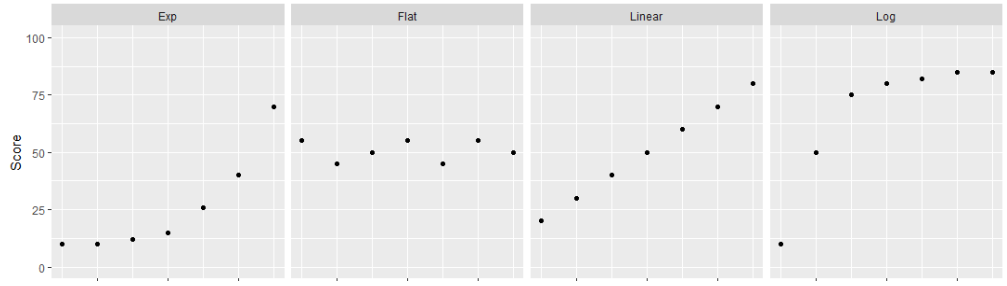
- **Location:** The Location will be using three levels: High, Middle and Low. The calculation of the Location is based on the median of the set of Finescores. The list of candidates consists out of 15 candidates, meaning that the median is the Finescore of the 8th candidate.

  All medians of the queryset will be split into three parts. Based on the boundaries of these parts, the querysets will be labeled to the High, Middle and Low levels.

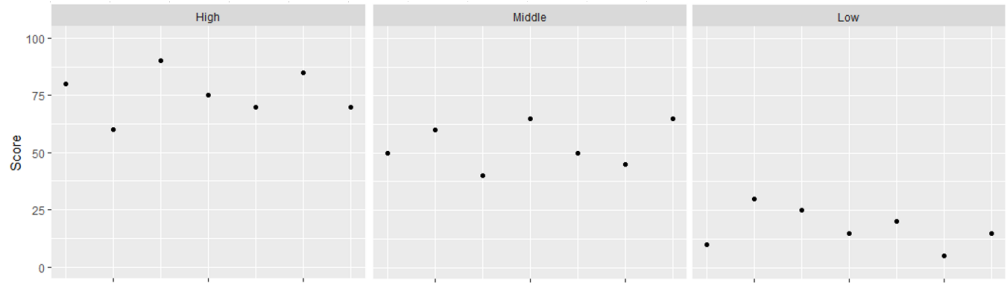
- **Spread:** The levels of the Spread factors are: High, Middle and Low, and are based on the standard deviation of the queryset. The standard deviation measures the dispersion of a queryset relative to its mean. The standard deviation is calculated as the square root of variance by:



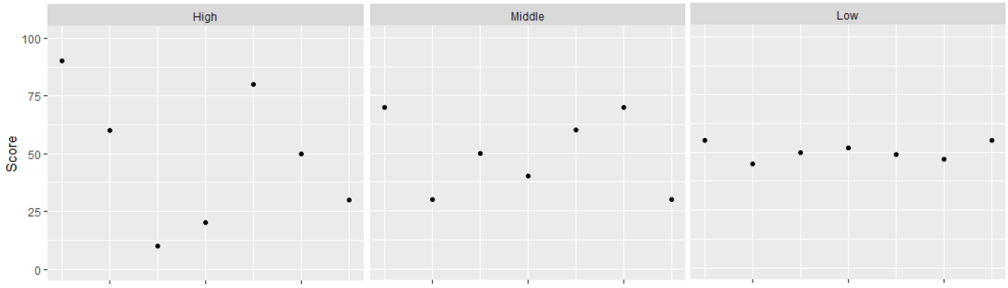
(a) Order



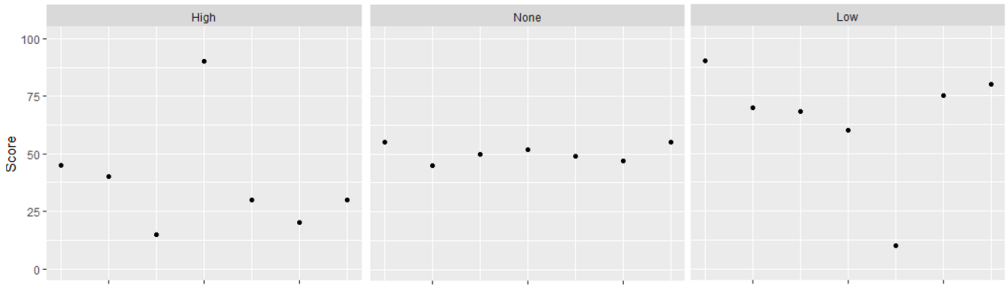
(b) Trend



(c) Location



(d) Spread



(e) Outlier

Figure 3.1: Visual illustration of the factor levels, with each level and example set of Fine scores.

Factor	Method	Equations	Threshold
Order	Spearman correlation	$\rho = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}$	H2L: $\rho < -0.2$ Random: $-0.2 < \rho < 0.2$ L2H: $0.2 < \rho$
Trend	Residual sums of squares	Linear: $y = b_0 + b_1 * x$ Exp: $y = x^b$ Log: $y = 1 - x^b$	Flat: $-2.2 < b_1 < 2.2$
Location	Median	8^{th} Finescore of list	High, Middle, Low
Spread	Standard deviation	$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$	High, Middle, Low
Outlier	Upper- and Lowerbound	High $> Q_3 + 1.5 * IQR$ Low $> Q_1 - 1.5 * IQR$	None: No outlier

Table 3.1: Overview of the thresholds for level labeling

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{3.5}$$

Where N is 15, the number of candidates. x_i the value of the i^{th} Finescore, and $\overline{x}$ the mean of the queryset. The same as Location, the standard deviation of all querysets will be split into three parts to make thresholds for the High, Middle and Low levels.

- **Outlier:** The levels of the Outlier Factor are: High, None and Low. The Outlier levels are calculated with:

$$\text{High} > Q_3 + 1.5 * IQR \tag{3.6a}$$

$$\text{Low} > Q_1 - 1.5 * IQR \tag{3.6b}$$

Where Q_1 and Q_3 are the first and third quartile of the set of Finescores, and IQR is the interquartile range. When all the Finescores of a set are between the given bounds, the queryset will be labeled with None.

3.3. Hypotheses

The hypotheses made on how the dependent variables are influenced by the independent variables are based on literature and intuition. This section will describe hypotheses about change in general, and about the distance.

3.3.1. Change

The hypotheses about how much annotators will change when a queryset belongs to certain factor levels are made for the main factors and some factor interactions which are the most interesting. The details for each hypothesis will be explained, for an overview, see table 3.2.

- **Order:** When the appearance of candidates to the annotator has a clear structure (H2L and L2H order), the annotator will be more confident about new judgements related to the previous judgements. Changes will occur when the annotator is not confident about the new judgement related to the previous judgements, which means that a clear structure in appearance of candidates will result in fewer changes.

When changes are made, they will be made at the beginning judgements in the opposite direction of the order. Meaning that H2L will have more positive changes and L2H more negative changes.

- **Trend:** Trend has 4 levels, with the flat trend corresponding, in most cases, to the random Order. This means that it has an unclear structure, resulting in more changes. The logarithmic and exponential trend have the characteristic of having a tail, which has most of judgements close to each other. The tail is at the end of a logarithmic trend, and at the beginning in an exponential trend. The part of judgements which is close to each other gives the annotator less confidence, meaning that more changes will be made.

- **Location:** Due to end aversion, annotators stay away from the extremes of the scale. When a new judgements is relatively closer to the end of the scales than previous judgements, the annotator will change the previous judgements. These changes are made in the direction to the middle, meaning that a high location has more negative changes, and a low location more positive changes.

- **Spread:** Having a low spread means that most of the judgements are close to each other, which limits the annotator to correct their previous judgements. A high spread gives the annotator space to adjust their new judgements to their previous judgement. However, when a change is made, the change will be bigger when the spread is higher.

- **Outlier:** An outlier in a set means that the other judgements are close to each other. When judgements are close to each other, more changes will be made. Concerning an outlier, the changes will be made in the direction of the outlier.

- **Order:Location:** In addition to the main effect explanation of Location, where annotators stay away from the extremes of the scale, the effect will be enlarged when the structure is towards that same edge. The most extreme interactions are H2L:Low and L2H:High. The direction of the change will be in the opposite direction of the closest edge.

- **Order:Spread:** From the main effect of Order, the hypothesis is that the clearer the structure, the fewer changes. For Spread, the higher the spread, the fewer changes are made. However, when changes are made due to the Order, the spread could influence the magnitude of the change. The absolute value of the change is probably higher when having a high Spread compared to a low Spread. In other words, the higher the spread, the lower the chance of a change, but when a change is made, the change will be bigger.

- **Trend:Spread:** The same hypothesis as Order:Spread, where the spread will influence the magnitude of the change. The Trend main effects still hold for this interaction.

3.3.2. Distance

When the over- or underrating is not corrected, and is a result of the queryset having certain properties, it is seen as a context effect. To see if there is an over- or underrating, the distance to the ground truth is taken. As discussed, change can be used as a metric to indicate that a potential context effect occurs. Annotators changing their judgements is not necessarily a problem as long as the judgements are close to the ground truth. When annotators change their previous judgements, it means that the annotator is aware of the over- or underestimation of the previous judgements. The annotator corrects the context effects by changing the judgements. When context effects occur which are not corrected, thus not changed, it means that the context effects have an impact on the set of judgements. Therefore, the hypothesis is: When changes are made, the context effects are corrected, which results in judgements closer to the ground truth.

Not only change can be used as a metric, but also the factor levels themselves. After the judgements are made, the set of judgements belongs to certain properties which are translated into the

Factor/Interaction	Change Hypothesis
Order	If there is a clear structure in the order of judgements, fewer changes will be made.
Trend	Judgements without a clear structure: flat trend and in the tail of a Log or Exp trend, have more changes.
Location	The closer judgements are to the ends of the scale, the more change will be made.
Spread	The higher the spread, the fewer changes will be made. However, when changes are made, the changes will be bigger in a high spread.
Outlier	If there is an outlier in the set, the other judgements are close to each other, which will result in more changes.
Order:Location	The closer the judgements are towards the ends of the scale, the more changes will be made. The most extreme cases: H2L:Low and L2H:High
Order:Spread	More changes are made in an unclear structure, these changes are bigger when the spread is higher.
Trend:Spread	Same principle as Order:Spread, changes are made in an unclear structure, the spread will influence the magnitude of the change.

Table 3.2: Factor hypothesis for the change behaviour of an annotator

factor levels. If the distance to the ground truth is different among certain factor levels, it is another indication of an over- or underestimation.

3.4. Data Collection

For this work, new data will be created and collected with the benefit of having control over how the music similarity judgements are collected and in which setting. Although there is control on how the data is collected, the process is as close as possible to MIREX to be able to make a fair comparison.

3.4.1. Instances Selection

The available AMS data of MIREX consists out of the tasks from the years 2006, 2007, 2009 till 2014. Not each year contains the same data concerning the amount of queries, candidates, and annotators, see table 2.1. The Finescores of the historical MIREX data can be used to create the new querysets. The Finescores are subjective and may be different among other annotators, however, they can be used as an indication because it can be assumed that the MIREX annotators are somewhere around the ground truth. The differences among annotators will be discussed in chapter 4, but for this chapter, the assumption is made that the MIREX annotator will be close enough to the ground truth to take the data as a useful reference.

The size of the querysets, i.e. amount of candidates, annotated with MIREX's E6K is dependent on which year the task is done. But for all years, the size is based on the output of all participating systems together. It is possible and happens frequently that different participating systems assign the same candidates as being similar. When this happens, the duplicates are removed such that judges are not assigning a score to the same candidates multiple times. The querysets in the MIREX data from 2006 till 2014 have varying sizes from 20 to 75 with two peaks around 25 and 60. These two peaks are due to the different amount of output candidates in certain years. The amount of candidates in the output of the system is dependent on what rules are applied per year, with the consequence of having varying sizes in the querysets. However, for this work the amount of candidates will deviate from the MIREX sizes for two reasons: 1) The judgements will be collected through crowdsourcing. In order to keep the amount of work reasonable for the crowdworkers, 15 candidates are chosen. This amount is the same amount used in a study about collecting music similarity judgement with crowdsourcing [28]. To give an example, when using 60 candidates in one queryset, it will take approximately 30 minutes for one

Factor	Number of Levels	Levels
Order	3	Decreasing (H2L), Random, Increasing (L2H)
Trend	2	Exponential (Exp), Logarithmic (Log)
Location	2	High, Low
Spread	2	High, Low
Outlier	3	High, None, Low

Table 3.3: Overview of the levels used for the full factorial design

Combination	Trend	Location	Spread	Outliers
1	Logarithmic	Low	High	None
2	Exponential	High	Low	Low
3	Exponential	High	High	Low
4	Exponential	High	High	High

Table 3.4: Non-existing combinations

judge, which is too much for a crowdworker; 2) The relation between amount of candidates and the specific Order factor has already been studied in the field of text information retrieval [22, 35]. They conclude that a minimum of 15 candidates is needed to show signs of an effect.

The new querysets with a fixed size of 15 candidates are extracted from the existing MIREX querysets. As already stated, these sets are larger and have to be reduced to get the required size of 15. This can be done by checking each permutation, with the benefit of increasing the chance to find instances for all conditions. However, this would result in a lot of permutations which requires too much computational power. For example, when having a MIREX set of 40, when reducing that set to 15 candidates it would have $5.26e+22$ permutations. Doing this for all queries would take too much time. Therefore, instead of all permutations, 500 samples of each queryset are made by randomly removing candidates till the size of the queryset is 15.

From all samples, queries have to be chosen such that the chosen queries in the dataset are varying enough. To do this, a full factorial design is applied to the factors. This means that all existing combinations of the levels across all factors are part of the dataset. The amount of queries needed to include all existing factor levels is $4^1 * 3^4 = 324$. However, due to the limited amount of resources, the amount of queries is reduced to 120. This is done by removing the middle levels of Location and Spread, and one by the changed labeling process of the Trend factor during the process of this work. Why this is changed will be further discussed in chapter 6. The querysets are chosen based on the most extreme levels, except for Order and Outlier. The most extreme levels are probably having a bigger magnitude when measuring context effects. An overview of the factor levels used to create the querysets can be found in table 3.3. This gives a total of $3^2 * 2^3 = 72$ combinations, however, out of the 72 combinations, 12 combinations are not found in all samples leaving the dataset with 60 queries. Without considering the order factor, the non-existing combinations can be found in table 3.4. The combinations in the table are mathematically difficult or not possible to create. Considering the very small chance of appearance of these combinations in a real-life scenario, it is not needed to make these combinations for the experiment.

During the process of the thesis, the amount of queries is doubled to 120 queries. All 120 querysets are different, but each of the 60 existing factor combinations will have 2 queries labeled to the combination.

3.4.2. Annotation Protocol

The created instances consist out $120 * 15 = 1800$ candidates, which have to be judged with crowdsourcing. To approach the crowdworkers, Amazon Mechanical Turk[1] is used. How this is done, and which measures are taken will be discussed in this section.

The collection of music similarity judgements will be similar to the collection method used in MIREX. In MIREX, the judgements are collected with the Evalutron 6000 (E6K) [19], which is developed for MIREX tasks involving human annotators. The E6K uses lists of candidates which are generated by the participating systems. These lists are, according to the systems, the most similar songs to a specific query. The annotators using the E6K are asked to give two similarity evaluations to each query-candidate pair: 1) a Broad category of similarity (i.e., Not Similar (NS), Somewhat Similar (SS), and Very Similar (VS)); and, 2) a Fine score between 0.0 (Least similar) and 10.0 (Most similar). The Fine score could also be between 0.0 (Least similar) and 100.0 (Most similar), depending on which year. For this thesis, only the Fine score is used.

Because the E6K is not available for usage in this work, the collection system is reproduced to an own web-based system. Same as with the E6K the annotator can listen to the query first, followed by the candidate, both query and candidate are 30 seconds snippets of the song. All candidates are listed in a row, with at the right side of the candidate player, a slider from 0 to 100. With this slider, the annotator can give their similarity score in the same way as the E6K. See figure 3.2 for a visual view for a query with the first two candidates.

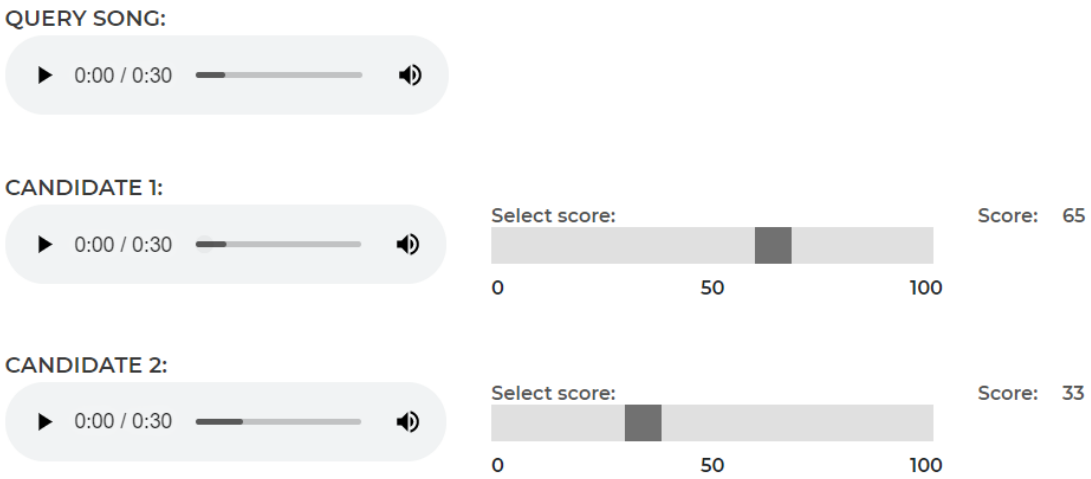


Figure 3.2: Visual representation example of how the annotators give their judgements.

During the session of the annotator, important actions are logged to the connected database. The following actions are logged:

- **Finescore:** When a score is given by the annotator, the Query ID, Candidate ID, Timestamp and the Finescore itself are logged.
- **Changed Finescore:** When a score is given to a candidate which has already been judged, it is also logged. This is considered as a change when another candidate is between the time of judging the initial and changed Finescore. Same as with Finescore, the Query ID, Candidate ID, Timestamp and the Finescore itself are logged.
- **Audio Play:** When a song is played, it will be logged with the Timestamp, QueryID and Candidate ID.

[1] `https://www.mturk.com/`

- **Audio Stop:** When a song is stopped, it will be logged with the Timestamp, QueryID and Candidate ID.

3.4.3. Quality Control

As already stated, the similarity judgements will be collected with crowdsourcing. To approach the annotators, Amazon Mechanical Turk will be used, this is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually[2]. As mentioned in section 2.5.1, crowdsourcing has some benefits compared to the expert used with MIREX, but also some disadvantages. The main argument for not using the crowd is that the crowd cannot always be trusted, the crowdworkers and their environment are not known, which makes it a not fully controlled experiment. However, studies using crowdsourcing have already shown some successful results, but each of them mentioned the importance of quality control.

The definition of quality is subjective and is depending on the requirements of the study to what level the quality measures will be taken. A general definition of quality is the "conformance to requirements" [9] or in other words "the extent to which the provided outcome fulfills the requirements of the requester" [1]. For the similarity judgements obtained with MTurk, the annotators need to fulfills the following requirements:

- Has a good reputation on the platform.
- Spent at least 5 minutes on the session.
- Listens to each song for at least 10 seconds.
- Effort is put into the session.

To check if the annotator fulfills the requirements, quality control measures are taken. The quality control measures ensure that the judgements are of a minimum level of quality. The quality control measures are:

- **Master Qualification:** MTurk is monitoring the tasks done by workers using their platform. Workers who have the master qualification are part of a specialized group of Workers who consistently demonstrate accuracy in performing a wide range of tasks. For using these workers, a 5% additional fee have to be paid. There is no need for additional implementation to make use of the master workers. MTurk is automatically filtering the master workers when asked for it.
- **Response time:** The whole session takes some time for an annotator. This includes reading the instructions, giving some basic information, listening to the songs, and thinking about which Finescore to give. Before submitting all tasks to MTurk, the task was tested on both MTurk and people not approached through MTurk. The average time spent in the test sessions was between 8 and 9 minutes, which is used as a base before submitting the tasks to MTurk. To give the crowdworkers some space, the requirement is that at least 5 minutes are spent on the session.

  This quality control measure can be applied by looking at the logs. For each action in the logs, a timestamp is given. By simply subtracting the timestamp at the start of the session with the timestamp at the end of the session, the total time is known. To ensure that an annotator is not taking a break such that enough time is spent, also the time between actions is considered.
- **Time spent on listening:** To get a proper understanding of the query and candidates, the annotator needs to listen to it for at least a significant part of the song. This is important to make a good comparison between the query and the candidates. Based on the test tasks, it can be seen that most of the annotators are not listening to the whole 30 seconds snippets of the song. It even is the case that most of the test annotators are only listening to half of the song snippets.

[2] `https://www.mturk.com/`

To ensure quality, the annotator needs to listen for at least 10 seconds to each song, which is taken as a threshold for this quality measure.

To apply this measure, the timestamps between the start of the played song is subtracted with the pause of the song. This is the time an annotator has listened to the song.

- **Query as a candidate:** To ensure that the judgements are made with effort and attention, the query itself is in the list of candidates. This means that the list of candidates is not of size 15, but 16. When the query is in the list of candidates, the annotator has to compare the query with the query itself. Comparing a song with the same song would mean that the Finescore should be 100, however, the test judgements showed that annotators are not completely confident about if it is exactly the same song and judge the song high but not 100. Probably because they paused or played the song at another timestamp, or they forget about how the exact query sounded like. To ensure that this common occurrence will not eliminate a lot of annotators, the Finescore of the query itself should be significantly high compared to the other judgements.

  To apply this to the data, the Finescore is taken where the Query ID is the same as the Candidate ID. This Finescore should be at least higher than the rest of the judgements by this annotator. The query as a candidate is not included in calculating the factor levels but will be added randomly to the list.

- **Compensation** It has been shown that financial incentives are important for crowdworkers [31]. The amount of the payment can be used to control tradeoffs among accuracy, speed, and total effort put into the task. With the resources available and ethical standards, the minimum wage of the United States is taken as a level of compensation.

  The payment process is done by MTurk. Because the judgements are given on an external web-based system, each annotator is given a unique ID which is linked to MTurk at the end of the session by the annotator. This ID can be used to approve or reject the work of the annotator.

The quality measures described are ensuring that the judgements have a certain quality which fulfills the requirements. It does not mean that the quality is comparable to the MIREX judgements. In chapter 4, the comparison between MIREX judgements and MTurk judgements will be compared.

3.5. Collected Data

The data which is collected with MTurk consists out of 120 querysets with each 16 candidates, where 15 candidates are real candidates and 1 candidate is the quality control measure. A total of 1920 judgements are collected in 3 different batches, each batch has 40 querysets.

The time needed to collect the judgements in significantly faster than the MIREX judgements, see table 3.5. Each batch was collected within 2 days, it is not known why the first batch was way faster than the others. However, taking into account the time needed for the MIREX judgements, some years 2 weeks, it is significantly fast. The most obvious explanation is that MTurk can run in parallel with a lot of annotators, while MIREX is restricted to a limited amount. The total cost of all jugements was 169,86 dollars including fees for MTurk.

A total of 153 querysets are submitted, where 33 (22%) are rejected and 120 (78%) are approved. When a queryset was rejected, it came available again for other annotators till the total approved amount of querysets was 120. The amount of rejections is lower than the AMS Broad score study by Lee [28], where 44.3% was rejected. Lee uses other quality control measures, which could explain the difference. Whether the approved querysets are reliable or not will be discussed in chapter 6. For an overview, see table 3.5.

3.5.1. Ground Truth

For each query-canditate pair, there is a total of 4 different Fine scores by 4 different annotators. The 4 annotators consist out of the MIREX, MTurk-H2L, MTurk-Random and MTurk-L2H annotator. For

Batch	Amount	Approved	Rejected	Time
1	54	40	14	16 Hours
2	49	40	9	26 Hours
3	50	40	10	28 Hours

Table 3.5: Amount of approved and rejected submitted querysets

each query-candidate pair, the annotators quite differ in scores given, also called the inter-rater agreement. To establish a ground truth or golden judgements, the average is taken over the other annotators. For a specific candidate judged by annotator i, the golden Fine score will be the average Fine scores given by the other annotators where $j \neq i$. In this way, the comparison to the ground truth will not include the Fine score of the annotator itself.

3.5.2. Factors

The new MTurk data is distributed and labeled in the same way as discussed in section 3.4.1. The MTurk querysets can deviate from the MIREX factor distribution, due to the subjectivity of different annotators. The factors will be calculated based on the golden judgements, which means that for each queryset the factors which are independent of the order of appearance to the annotator are likely to be the same. These factors are Location, Spread and Outlier. Due to the golden judgements being calculated over the Fine scores of the other annotators, the Location, Spread and Outlier can still be different.

4

Reliability MTurk Judgements

As described in section 3.4.2, the similarity judgements are obtained using the crowdsourcing platform Mechanical Turk[1]. In comparison to the MIREX experts, crowdsourcing is beneficial considering the cost and time [4, 39]. However, when using an unknown crowd of people, there is a risk of lower quality work. Although quality control measures are taken, these measures will not by definition eliminate the difference between an expert and a non-expert. The quality control measures are filtering the people who did not have attention or did not put effort into the list of judgements, this does not automatically mean that the Fine scores of a non-expert are comparable to the ones of an expert. Therefore, the MTurk judgements have to be compared to the MIREX judgements to see if they are reliable and can be used for further analysis.

Music similarity judgements are known to be subjective, meaning that there could be differences across Fine scores between multiple annotators. The AMS task has two different scores, the Broad score and the Fine score. To see how big the difference between annotators is, in other words, how much they agree on the Fine score of the same query-candidate pair, the inter-rater agreement can be used. For both Broad and Fine score, the inter-rater agreement has been studied [17, 25]. This was only possible for the 2006 edition, which had three different annotators for each query-candidate pair. Both studies are based on judgements obtained with the Evalutron6K [19]. Music similarity judgements obtained with crowdsourcing in the current annotation protocol have only be studied for the Broad score [28], not for the Fine score, which is a gap to be filled 4.1. Therefore, this chapter will analyse the gap of crowdsourced Fine scores by answering the question if music similarity judgements obtained through crowdsourcing are reliable.

4.1. Data

The data used for analysing the reliability consists out of three parts:

- MTurk data: the data which have to be analysed for reliability. The data consists out of three annotators for each query-candidate pair. Each queryset has 15 candidates.

[1] `https://www.mturk.com/`

	Broad score	Fine score
MIREX	[25]	[17]
MTurk	[28]	This Thesis

Table 4.1: Studies about agreement between annotators

	Graders per Queryset	Total Querysets	Size of Queryset	Total judgements
MTurk	3	40	15	1784
MIREX 2007-2014	1	40	15	600
MIREX 2006	3	60	30	5400

Table 4.2: Data used for analysing MTurk reliability

- MIREX 2006-2014 data: the MTurk querysets are created based on MIREX data. Each MTurk queryset of 15 candidates also has a set of the same 15 candidates in the MIREX data, which already has a Fine score given by a MIREX annotator.
- MIREX 2006 data: the AMS task in the 2006 edition consists out of three annotators for each query-candidate pair. Querysets differ in size from 30 to 60 candidates.

See table 4.2 for more detailed statistics for each dataset. To answer the question if the MTurk data is reliable, the comparison will be made in two ways: 1) Comparing the inter-rater agreement of pairs of annotators in the MTurk and 2006 data separately from each other, and 2) Comparing the MTurk data with the corresponding MIREX data.

4.2. Inter-rater Agreement

The inter-rater agreement measures the agreement between different annotators. There are multiple ways of measuring the inter-rater agreement, in this section, the inter-rater agreement will be measured by the correlation between pairs of annotators, the error between pairs of annotators and by looking at the agreement of pairs of Fine scores.

4.2.1. Correlation

For the Broadscore it has been shown that there is a "fair" level of agreement [25]. They studied the correlation for both the original three-level Broad scores, but also a two-level Broad score, by concatenating the Very Similar Broad score with the Somewhat Similar Broad score. The level of agreement is based on Fleiss's Kappa score which is a measure of inter-grader reliability for nominal data, and is based on Cohen's two-grader reliability Kappa, but measures reliability among an arbitrary number of graders [16]. The Fleiss's Kappa of the three-level and two-level Broad score is 0.2141 and 0.2989. Considering the range of a fair level of agreement is between 0.21 – 0.40, the inter-rater agreement of the Broad score is at the low end of the range.

The Broad score has a "fair" level of agreement, but this is based on three and two-level categorical scores. The numeric Fine score is different due to the wider range of values the annotator can choose from. Therefore, the expectation would, intuitively, be that annotators will have a lower correlation with each other, compared to the Broad score. Fleiss's Kappa was used to measure the agreement with the Broad score, but this score is only suitable for categorical variables, for the Fine score other correlation measures are used to show the inter-rater agreement between annotators. Flexer showed that the Pearson correlation between pairs of annotators for all candidates together on average ranges from 0.37 to 0.43 in the 2006 data [17].

To answer the question if the MTurk data is reliable, a comparison has to be made with the 2006 data. In terms of correlation, the comparison will be made by comparing the correlation of the pairs of annotators in the 2006 data versus the pairs of annotators in the MTurk data, but also by comparing the 2006 correlations with the MTurk correlations. The correlation will be measured with the Pearson correlation and the Spearman correlation. The Pearson correlation is used to evaluate the linear relationship between two Fine scores and therefore taking proportions into account. It is calculated with the equation from 3.1. Spearman correlation is calculated with the following formula:

$$\rho = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{4.1}$$

Pearson	2006	MTurk
Mean	0.27	0.36
Median	0.30	0.46
Min	-0.51	-0.61
Max	0.83	0.89
SD	0.30	0.36

(a) Pearson correlation

Spearman	2006	MTurk
Mean	0.23	0.33
Median	0.27	0.40
Min	-0.52	-0.62
Max	0.83	0.90
SD	0.29	0.37

(b) Spearman correlation

Table 4.3: Correlation statistics for each queryset between all pairs of judges for the 2006 and MTurk data. Both datasets has three different judges for each queryset.

where rg_X and rg_Y are the X and Y converted to ranks, cov is the covariance and $\sigma_{rg_X}\sigma_{rg_Y}$ are the standard deviation of rg_X and rg_Y. Both correlation measures are in essence similar, but Spearman uses the rank for each candidate in the queryset instead of the Fine score itself. It evaluates in a monotonic way and therefore somewhat normalizes the Fine scores of the annotators.

The correlation values of the mean, median, minimum, maximum and standard deviation are shown in table 4.3. For the MIREX data, there is a small difference in correlation between Pearson and Spearman. Because Spearman evaluates monotonically, it was expected that this would result in higher values. However, this is not the case in the data. This is probably the case due to outliers in the candidate sets, which influences the Pearson correlation, but not the Spearman correlation. In both cases, the median is higher than the mean, because of the logarithmic shape when putting all correlation values together in a trend. The values for both correlations are quite low, this is especially interesting taking into account that the judgements are made by experts. This is confirming the subjective nature of music similarity judgements described by Flexer [17]. The same correlations are calculated for the MTurk data, in table 4.3 it can be seen that the correlation between pairs of annotators from MTurk is higher than the pairs of annotators from MIREX. This was not expected and could be interesting considering time and costs. Figure 4.1 shows a more detailed view of the values in table 4.3.

It has to be noted that the sizes of the querysets are different between MIREX and MTurk, with MIREX having 30 candidates and MTurk 15 candidates. When calculating correlation, the size of the sets can influence the correlation in terms of confidence [5]. To see what happens with the correlation values when the MIREX queryset also has 15 candidates instead of 30, 5 random samples of size 15 are created for each MIREX queryset. The correlation between pairs of annotators for all samples is calculated in the same way as described before. Results are shown in table 4.4, the correlation values are highly similar. Therefore the original MIREX queryset size will be used for further analysis.

Overall, the MTurk data seems reliable in terms of correlation. The correlation between pairs of graders in the MIREX data is lower than pairs of graders in the MTurk data. Although the standard deviation is higher in the MTurk data, the MTurk data is reliable when using sufficiently large enough datasets.

The MTurk data is based on MIREX Fine scores, so the agreement can not only be measured between MTurk annotators, but also together with their corresponding MIREX Fine scores. Each query-candidate pair has 3 MTurk annotators and 1 MIREX annotator. The correlations will again be made with the Pearson 3.1 and Spearman 4.1 correlation. The Pearson and Spearman values seems highly comparable to previous tables, so for the remainder of this section, only the Pearson correlation will be used.

Lee studied the correlation between MTurk Broad scores and MIREX Broad scores, a Pearson correlation of 0.495 was found [28]. This correlation is based on the whole dataset, not for each queryset individually. When taking the whole set of MTurk Fine scores and their corresponding MIREX Fine scores, the Pearson correlation is 0.401. This is slightly lower than the Broad score Pearson corre-

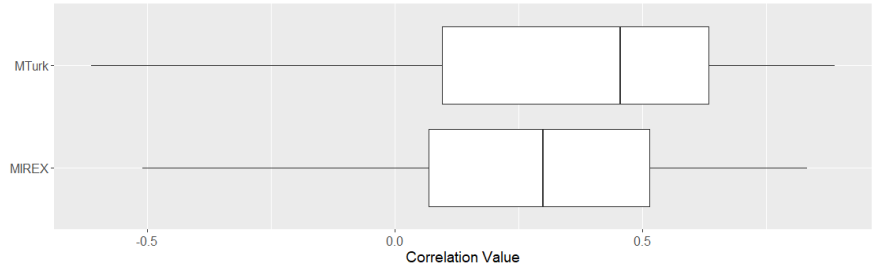
Pearson	2006	2006 samples
Mean	0.27	0.26
Median	0.30	0.30
Min	-0.51	-0.51
Max	0.83	0.78
SD	0.30	0.30

(a) Pearson correlation

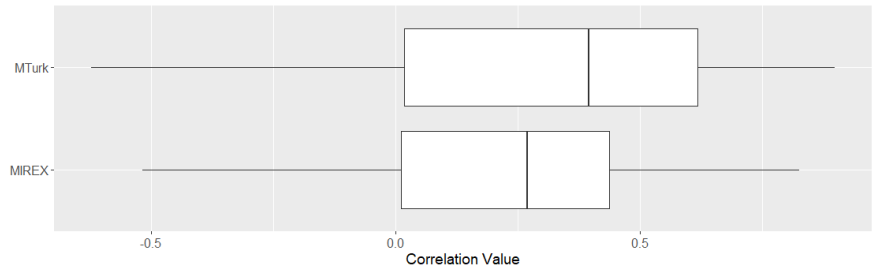
Spearman	2006	2006 samples
Mean	0.23	0.22
Median	0.27	0.28
Min	-0.52	-0.55
Max	0.83	0.76
SD	0.29	0.29

(b) Spearman correlation

Table 4.4: Correlation statistics between all pairs of judges for the 2006 querysets of size 30 and the random samples of size 15. Both datasets has three different judges for each queryset.



(a) Pearson correlation values



(b) Spearman correlation values

Figure 4.1: Boxplots of the correlation values, with on each figure, the MIREX set on the left and the MTurk set on the right. For both Pearson and Spearman.

	MTurk-MIREX
Mean	0.41
Median	0.48
Min	-0.54
Max	0.97
SD	0.37

(a) Pearson correlation

	H2L-MIREX	L2H-MIREX	Random-MIREX
Mean	0.45	0.41	0.37
Median	0.56	0.47	0.46
Min	-0.40	-0.54	-0.44
Max	0.92	0.97	0.85
SD	0.32	0.38	0.39

(b) Pearson correlation

Table 4.5: The correlation between all MTurk annotators and the MIREX annotator (a), and the MTurk annotators separated in Order for the Pearson correlation.

lation, however, this could be explainable due to the difference in the three-level categorical Broad score and the continuous Fine score. Both values are comparable to other studies related to experts versus MTurk annotators [40].

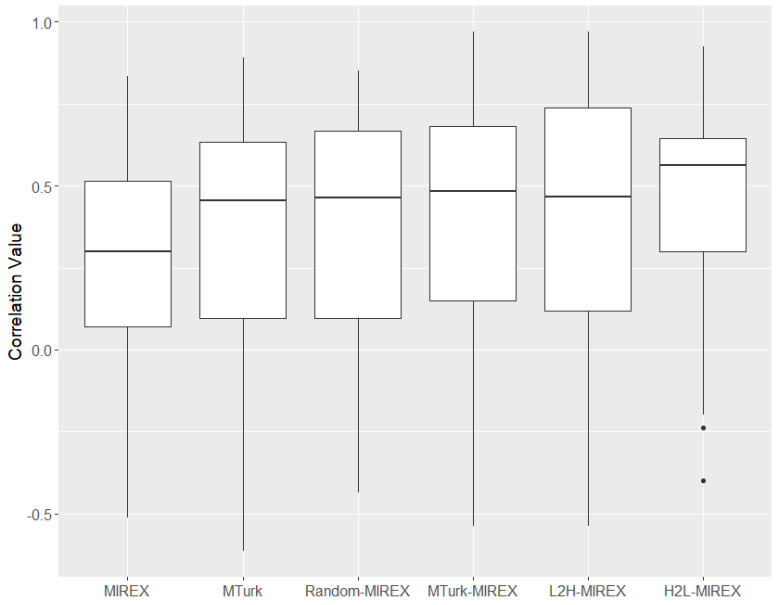
The correlation values are different when calculating the correlation value for each queryset individually, where each MTurk queryset is compared with the corresponding MIREX queryset. Table 4.5a shows the Pearson correlation for all MTurk annotators compared with their MIREX annotator. The values are higher than the correlations between annotators in MIREX and MTurk separately. Table 4.5b shows the correlation for each of the three MTurk annotators separately compared to the corresponding MIREX set, as described in section 3.2, each annotator has a different order, which could influence the correlation in a positive way. This will be further analysed in chapter 5. A more detailed view of the values in table 4.5, can be seen in figure 4.5, the correlations between pairs of annotators in the MIREX and MTurk data separately are shown as a reference.

Overall, the correlation values seem to be even higher when calculated between the MTurk annotator and the original MIREX annotator. It was already shown that the correlation between MTurk annotators pairs was higher than the pairs of the MIREX data, although the difference is not that much. Therefore, the MTurk judgements are comparable with the MIREX judgements in terms of correlation.

4.2.2. RMSE

The values of the correlation in the previous section are based on the whole set of Fine scores in the queryset, it is representing the linear relation with the Pearson correlation based on the best fitted linear line. However, the Pearson values shown in the previous section are not showing how far the Fine scores are from the line, i.e. what the error is. For comparing the error, the root-mean-square error will be used in this section. For this section, the 2006 and MTurk data will be used, which both have three different annotators. The MIREX data has a total of 180 querysets, while the MTurk data has 120 querysets.

For the calculation of the RMSE, the list of Fine scores of one annotator is used together with the average of the corresponding Fine scores of the other two annotators. The RMSE is calculated between these two lists, note that this is slightly different than the ground truth as described in 3.5.1. The RMSE values can be found in figure 4.3, with the general statistics in table 4.6. The RMSE in the 2006 data seems to be lower than the MTurk data. This would mean that the error between an annotator and the average of the other annotators is smaller in the 2006 data than the MTurk data. A possible explanation would be that the MTurk data consists for a big part out of extreme querysets, which are more prone to context effects, this will be further discussed in chapter 5. The three different annotators in the MTurk data have annotated the querysets in H2L, Random and L2H Order, while the MIREX data is judged in Random order by all annotators. When a queryset is annotated by an annotator who was not aware of the effect, the Fine scores will be further away from the other annotators, which could possible explain the higher RMSE in combination with the higher correlation.



(a) Pearson correlation values

Figure 4.2: Boxplots of the Pearson correlation values between each MTurk annotator (indicated with an Order) and the MIREX annotator. The correlation between annotators from Figure 4.1 are given as reference. The boxplots are sorted on increasing order.

RMSE	2006	MTurk
Mean	26.09	29.21
Median	25.79	29.22
Min	11.33	12.96
Max	46.06	50.70
SD	6.97	8.89

Table 4.6: RMSE statistics for the 2006 and MIREX data.

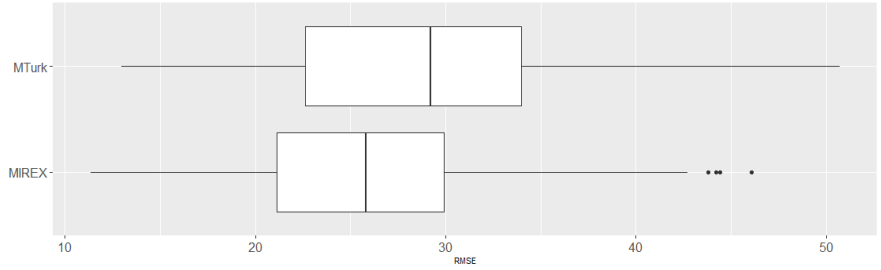


Figure 4.3: RMSE values per queryset between Fine scores of one annotator and the average Fine score of the other two annotators. Both for the MIREX and MTurk annotators.

4.2.3. Pairwise Fine Score

The correlation in the previous section is based on the whole set of Fine scores in a queryset, this section will not analyse the whole queryset, but each Fine score individually. Each query-candidate pair has three different MTurk annotators, the Fine scores of these three annotators are compared. The 2006 MIREX edition has also three different annotators for each query-candidate pair, which can be used as a baseline to check the reliability of the MTurk data.

The distribution of the MTurk Fine scores is shown in Figure 4.4. H2L, Random and L2H indicate the different annotators from MTurk, the corresponding MIREX Fine score distribution is also shown. Each Fine score is added to a certain interval [0, 10), [10, 20) to [90, 100], indicating the interval to which the Fine score belongs. When exploring the distribution for each interval, they are quite similar among the different annotators. The distribution on its own is not a solid measure, but is a first indication that the Fine scores are distributed over the same intervals. Note that the Fine scores are not equally distributed over all intervals, due to created querysets with factor Location having levels High and Low. Middle is not part of the creating process which explains the absence of the Fine scores around the middle.

For analysing each query-candidate pair individually, the absolute distance between the Fine scores of three different annotators is taken. To measure this distance, the Fine score of annotator i is taken and compared to the average of the Fine scores of the other two annotators $j \neq i$. Flexer showed this for the MIREX 2006 data, for example, the data showed that Fine scores in the (90, 100] interval, is given an average Fine score of 65.4 by the other graders [17]. All MIREX and MTurk intervals are shown in figure 4.5a, the dashed diagonal line indicates the perfect agreement. When taking the value of the Fine score itself instead of the interval, see figure 4.6, it can be seen that in both MIREX and MTurk the points are following the trend, but with a very high range of values on the scale. When comparing the results for MIREX and MTurk, it can be seen that the lines are almost similar. MTurk has a slightly lower avg Fine score at last intervals, but considering the correlation values, this difference can be neglected. When splitting the MTurk data into the different orders, fig. 4.5b, the other annotators have a slightly higher average when the interval annotator has a H2L order. But again, this can be neglected taking into account the correlation.

The comparison of the pairwise Fine scores between MIREX and MTurk are quite similar, but it is clear that agreement between annotators for both MIREX and MTurk is rather low, which is in line with the already shown correlation values.

4.3. Summary

In this chapter, the collected data through MTurk is discussed to compare the data with the MIREX data. The MTurk data is collected according to the same annotation protocol as the MIREX data, with the difference of having non-experts as annotators. To check the reliability of the MTurk annotators, the data is compared with the MIREX data in two ways, 1) Comparing the inter-rater agreement of pairs of annotators in the MTurk and 2006 data separately from each other, and 2) Comparing the MTurk data with the corresponding MIREX data.

The inter-rater agreement is measured for the 2006 data and the MTurk data separately, both datasets have three different annotators. The inter-rater agreement is measured with the correlation between annotators, the RMSE between annotators and with pairwise Fine score comparisons. The correlation is calculated with Pearson and Spearman, both measures are comparable, so Pearson is used as a base. The correlation values show promising results, the MTurk data has even a better correlation between annotators than the MIREX data. It has to be noted that this can be due to the Order in which the MTurk annotators are giving their judgements, but this will further be analysed in chapter 5. Overall, the MTurk correlation is highly comparable to the MIREX data. Although the higher correlation values for MTurk, the MTurk data has a bigger error concerning the RMSE. For each annotator, the RMSE is calculated with the list of Fine scores of the annotator together with the average of the corresponding Fine scores of the other two annotators. The RMSE between annotators in the MTurk data is higher than the 2006 data, meaning the Fine scores of MTurk annotators are further away from each other. Again, this could be due to the Order in which the MTurk annotators are giving their judge-

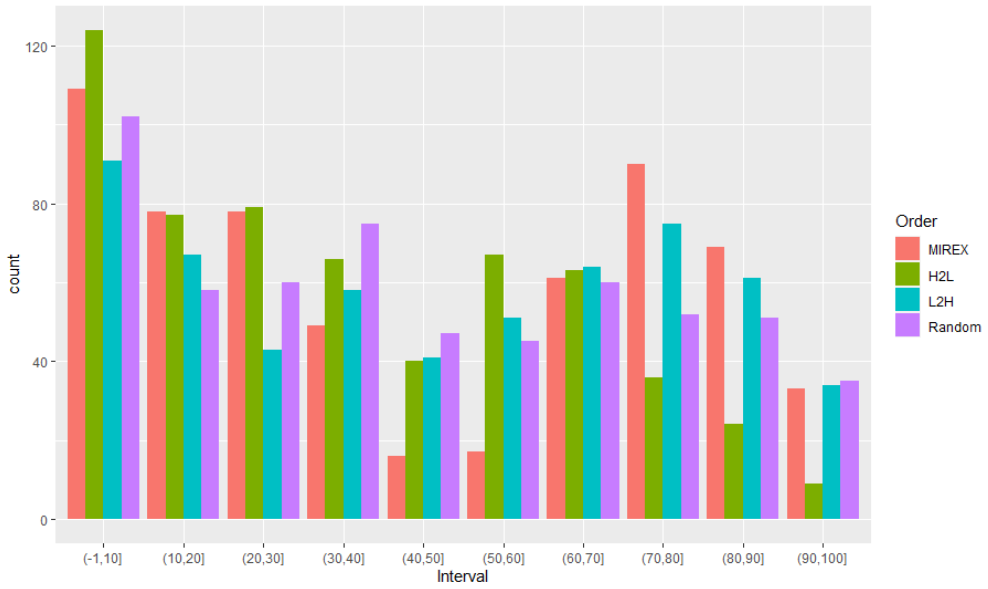
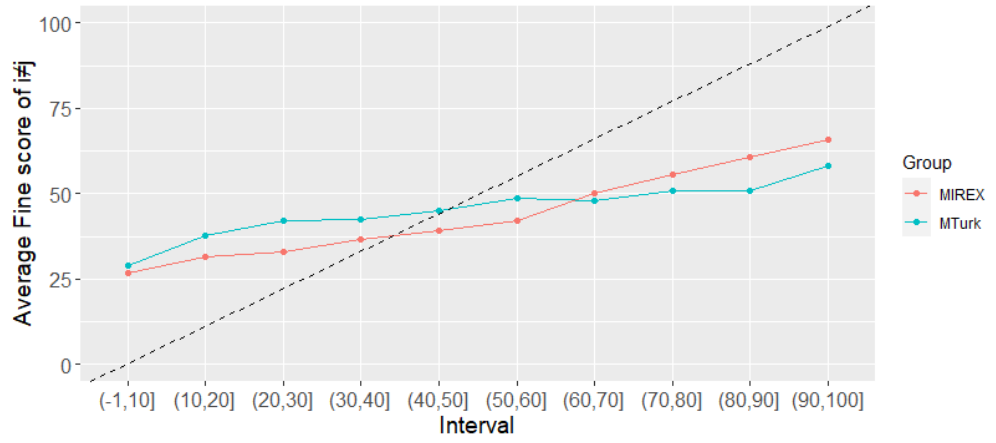


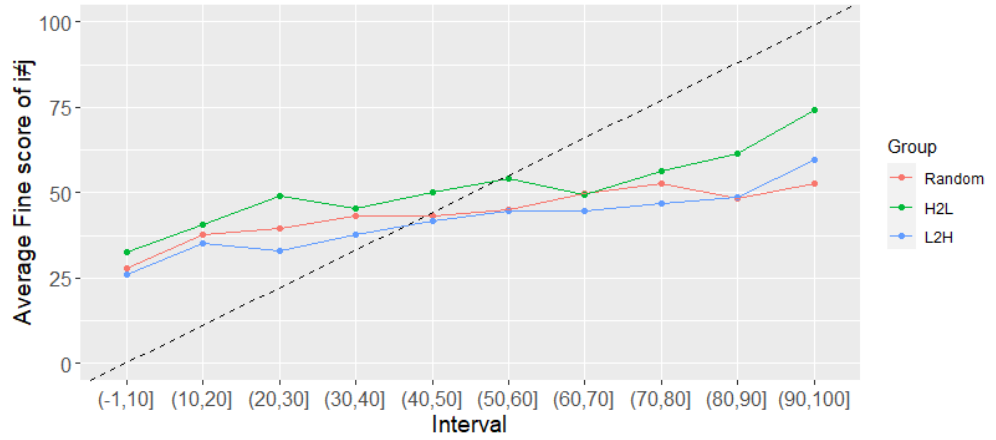
Figure 4.4: Distribution of all Fine scores for both MTurk and MIREX. The scale is divided into 10 intervals.

ments. However, the difference in RMSE is not of such magnitude that the data seems unreliable. When looking at the difference between Fine scores for each query-candidate pair individually, the MIREX and MTurk data are highly comparable. This is done by taking the absolute distance between the Fine scores of three different annotators. To measure this distance, the Fine score of annotator i is taken and compared to the average of the Fine scores of the other two annotators $j \neq i$.

The MTurk data is based on the MIREX data and thus has corresponding MIREX query-candidate pairs. When comparing the MTurk data with their corresponding MIREX data, the correlation is higher than between annotators in both the MTurk and MIREX data. The reason for this is not explored, but it is in favor of the reliability of the MTurk data. Overall, the MTurk data is comparable with the MIREX data and will therefore be used for analysing the context effects.



(a) MIREX 2006 and MTurk



(b) MTurk orders

Figure 4.5: For each judgement i in both MIREX as MTurk (a), the average of the other two graders $j \neq i$ is given. i is shown in the interval, the average is shown as a line. (b) shows the different annotators in MTurk.

Figure 4.6: For each judgement i in both MIREX (a) and MTurk (b), the average of the other two graders $j \neq i$ is given.

5

Context Effects in Music Similarity Judgements

This chapter will address the two research questions by 1) recognizing context effects and 2) measuring the impact of the context effect. To answer the first part, a metric is needed to see which querysets are prone to context effect. The metric will be the amount of changes made by the annotator. In a judgement session of an annotator, some will notice an effect, some will not. When an effect is noticed by the annotator, they will change their previous judgements and therefore adjust their previous over- or underestimation of the judgements. When not being noticed, this over- or underestimation is not corrected and therefore has some distance to the true value of the judgement. This distance is used to answer the second part, the impact of the context effects.

5.1. Change

The logs in the MTurk data shows that around half of the annotators go back and change at least one of the candidates to a new Fine score. The change is used as a metric to indicate if a context effect exists. Hypotheses are made about the factors as independent variables and the change as dependant variable. The hypotheses are made for the main factors and some factor interactions which are the most interesting. The hypotheses are already discussed in section 3.3, where the main points are:

- **Order:** If there is a clear structure in the order of judgements, fewer changes will be made.
- **Trend:** Judgements without a clear structure: flat trend and in the tail of a Log or Exp trend, have more changes.
- **Location:** The closer judgements are to the ends of the scale, the more change will be made.
- **Spread:** The higher the spread, the fewer changes will be made. However, when changes are made, the changes will be bigger in a high spread.
- **Outlier:** If there is an outlier in the set, the other judgements are close to each other, which will result in more changes.
- **Order:Location:** The closer the judgements are towards the ends of the scale, the more changes will be made. The most extreme cases: H2L:Low and L2H:High
- **Order:Spread:** More changes are made in an unclear structure, these changes are bigger when the spread is higher.
- **Trend:Spread:** Same principle as Order:Spread, changes are made in an unclear structure, the spread will influence the magnitude of the change.

Response Variable	Description
Count	The total amount of changes made in the queryset
Total	The absolute value of all changes in a queryset together
Average Total	The absolute value of all changes in a queryset together divided by the amount of changes made in the queryset. Total change divided by Count
Direction	The total value of positive changes subtracted by the total value of negative changes in a queryset
Average Direction	The total value of positive changes subtracted by the total value of negative changes in a queryset divided by the amount of changes made in the queryset. Direction of change divided by Count
Where	Each candidate number is given a rank between 0-1 based order of appearance. The average rank of all changed judgements in a queryset is taken to indicate where the changes are made

Table 5.1: Change response variables

Factor	Count	Total	AvgTotal	Direction	AvgDirection	Where
Order	Random >H2L Random >L2H	Random >H2L Random >L2H	Random >H2L Random >L2H	Random <H2L Random >L2H	Random <H2L Random >L2H	Random >H2L Random >L2H
Trend	Flat >Exp Flat >Log Exp >Linear Log >Linear	Flat >Exp Flat >Log Exp >Linear Log >Linear	Flat >Exp Flat >Log Exp >Linear Log >Linear			Exp >Flat Exp >Linear Log >Flat Log >Linear
Location	High >Middle Low >Middle	High >Middle Low >Middle	High >Middle Low >Middle	High <Middle Middle <Low	High <Middle Middle <Low	
Spread	High <Middle Middle <Low		High >Middle Middle >Low			
Outlier	High >Middle Low >Middle	High >Middle Low >Middle	High >Middle Low >Middle	High >Middle Middle >Low	High >Middle Middle >Low	

Table 5.2: Hypotheses for each response variable, displayed in contrasts for the applicable factor levels.

Not all hypotheses can be captures based on one value of change, therefore six response variables are created to capture the different ways of looking at change. The response variables are Count, Total change, Average Total change, Direction of change, Average Direction of change and Where the change is made, for an explanation, see table 5.1.

To see if the MTurk data is in line with the hypotheses, comparisons have to be made between the factor levels. For example, the Order hypothesis stated that the clearer the structure, the fewer changes will be made. This means that the H2L and L2H order has fewer changes than the Random order. When applicable to the response variable, pairwise comparisons can be made for all the levels in a factor or factor interaction. Table 5.2 shows the comparison hypotheses for each applicable pairwise factor level.

5.1.1. Analysis Method

To see if the data is in line with the hypotheses, the lmer[1] package in R will be used for fitting the models. The package provides functions to fit and analyze linear mixed models. The Linear Mixed-Effects Models are used due to the kind of messy data. First of all, the model has to deal with a lot of parameters. The factor levels are taken as independent variables in the model, meaning that there are a lot of combinations with the data available. The consequence of having a lot of combinations is that the sample size per combination is relatively small. On top of that, the factor levels are not truly independent. The lmer models deal with these kinds of problems.

From the model, the contrasts between pairwise factor levels are calculated. The contrasts are based on the estimated marginal means (EMM), which are the means for the factors averaged for all levels of the other factors, which will result in taking care of the unbalanced data. When taking the normal mean, the results will be biased due to the imbalance of the dataset. Depending on the model, some contrasts do not have enough data, and thus combinations, to calculate the significance of the contrast. This is especially the case when the model uses the whole set of factors with all two-level interactions. In this case, it is complex to calculate the significance of the contrast. Therefore, two different ways of evidence are used: soft and hard evidence.

5.1.2. Soft Evidence

With the soft evidence, all factors are used in the model as main effects, together with their two-level interactions. As already described, some combination do not exist which makes it difficult to calculate the estimated marginal means. However, the reference grid of the model includes predictions of all the combinations which do exist. These combinations can be used to calculate the mean of the existing predictions, which will be used as the estimated marginal mean. This is not the fully correct way of calculating estimated marginal means, but it gives an indication based on the existing means. There is not enough statistical power to see if any significant difference exists between certain factor levels, but it could indicate if the data is in line with the hypotheses. This way of looking at how much the data is in line with the hypotheses will be used as soft evidence.

For each change response variable, a model is made. For all hypotheses, the corresponding marginal means and contrasts are calculated and showed in 5.3. The table cells are colored according to if the contrast is in line with the hypotheses. Green indicates that the contrast is in line, red indicates that the contrast is not in line, and white means that the response variable or contrast is not applicable to the hypotheses.

Table 5.4 shows for each factor the percentage of how many contrasts are in line with the hypotheses. A red cell means below 25%, yellow between 25% and 75% and green cell is above 75%.

5.1.3. Hard Evidence

For the hard evidence, not the full model used for soft evidence is taken, but the model with significant results. To obtain a model which can be used for hard evidence, the full model, including main effects and two-level interactions, is taken as starting point. Iterations will be made over this model, where the least significant interaction is removed till only main effects with significant interactions are left. The least significant interaction is removed based on Analysis of variance (Anova), which looks at the variation of the groups and where the variation is. This is done by looking at the variation between groups and compare it with the variation within the groups. When the least significant contrasts are removed, the significant contrasts will be used as hard evidence. For the significance, the 0.05 p-value threshold is taken.

The hard evidence is showed in 3.2 by a darker green or red color. The table is only showing significant contrasts which are part of the hypotheses. There are a few observations which are not part of the hypotheses, but show significant differences among contrasts. However, only three significant observations are outside the hypotheses which are not worth mentioning or used for further analysis

[1] `https://cran.r-project.org/web/packages/lme4/`

Factor	Level Contrast	Count	Total	AvgTotal	Direction	AvgDirection	Where
Order	H2L - Random	-0.14	23.22	5.01	-13.79	-2.03	-0.26
	H2L - L2H	-1.77	-7.61	-5.52	-10.31	-5.30	0.41
	Random - L2H	-1.63	-30.83	-10.52	3.48	-3.27	0.67
Trend	Exp - Linear	-2.24	-53.55	-9.94	8.63	-4.78	-0.48
	Exp - Flat	-3.15	-59.93	-15.40	-21.53	-11.10	-0.16
	Exp - Log	-2.62	-37.41	-10.38	-34.15	-12.62	0.57
	Linear - Flat	-0.91	-6.38	-5.45	-30.16	-6.33	0.32
	Linear - Log	-0.38	16.15	-0.44	-42.78	-7.84	1.05
	Flat - Log	0.53	22.52	5.02	-12.62	-1.52	0.73
Location	High - Middle	-1.98	-11.64	-3.47	9.13	0.32	0.55
	High - Low	-5.13	-9.79	0.68	-15.37	-6.73	0.54
	Middle - Low	-3.15	1.85	4.15	-24.49	-7.05	-0.00
Spread	High - Middle	0.38	5.62	-6.12	-3.73	-1.47	0.50
	High - Low	-0.25	-20.75	-9.88	17.55	7.05	-0.10
	Middle - Low	-0.63	-26.37	-3.76	21.28	8.51	-0.61
Outlier	High - None	-0.58	7.59	8.77	26.64	14.41	-0.32
	High - Low	-2.39	-23.38	9.11	28.40	15.82	-0.38
	None - Low	-1.81	-30.97	0.35	1.76	1.41	-0.06
Order:Location	H2L(High - Middle)	-2.16	-15.66	-9.73	13.45	0.04	0.72
	H2L(High - Low)	-5.37	-28.83	1.56	1.49	-5.73	0.53
	H2L(Middle - Low)	-3.21	-13.17	11.29	-11.96	-5.77	-0.19
	Random(High - Middle)	-3.02	-3.49	-2.76	13.92	-7.93	0.83
	Random(High - Low)	-6.95	-8.98	1.06	8.13	-7.41	0.53
	Random(Middle - Low)	-3.93	-5.49	3.82	-5.79	0.53	-0.30
	L2H(High - Middle)	-1.07	-12.16	3.70	0.65	6.17	0.39
	L2H(High - Low)	-3.49	7.40	1.07	-44.75	-6.25	0.78
	L2H(Middle - Low)	-2.42	19.56	-2.64	-45.40	-12.42	0.38
Order:Spread	H2L(High - Middle)	0.80	14.54	-11.79	-4.77	-1.71	0.61
	H2L(High - Low)	1.18	-8.86	-14.99	25.45	5.93	-0.60
	H2L(Middle - Low)	0.38	-23.40	-3.20	30.22	7.64	-1.21
	Random(High - Middle)	-1.34	-16.25	-20.21	-20.03	-9.48	0.78
	Random(High - Low)	-1.73	-13.14	-18.80	-2.29	1.56	0.44
	Random(Middle - Low)	-0.39	3.11	1.40	17.74	11.05	-0.34
	L2H(High - Middle)	1.06	10.40	9.17	8.07	4.04	0.27
	L2H(High - Low)	-0.81	-38.46	1.42	22.12	11.83	0.15
	L2H(Middle - Low)	-1.87	-48.86	-7.75	14.05	7.79	-0.11
Trend:Spread	Exp(High - Middle)	3.10	53.85	20.70	13.34	10.16	
	Exp(High - Low)	1.13	15.86	2.21	13.42	6.46	
	Exp(Middle - Low)	-1.97	-37.98	-18.50	0.08	-3.70	
	Linear(High - Middle)	0.42	-17.89	-0.29	-2.64	4.88	0.05
	Linear(High - Low)	-1.06	-59.99	-3.34	29.76	6.74	-0.47
	Linear(Middle - Low)	-1.48	-42.10	-3.05	32.40	1.86	-0.52
	Flat(High - Middle)	1.56	27.99	-4.42	3.12	-7.25	0.94
	Flat(High - Low)	0.77	-1.90	0.81	14.89	5.78	0.29
	Flat(Middle - Low)	-0.79	-29.88	5.23	11.77	13.02	-0.64
	Log(High - Middle)	-1.53	-13.59	-18.66	-15.73	-2.93	0.38
	Log(High - Low)	-1.09	-22.94	-28.07	13.15	8.67	-0.26
	Log(Middle - Low)	0.45	-9.35	-9.41	28.88	11.61	-0.65

Table 5.3: Contrast of the levels for each change response variable.

Factor	Count	Total	AvgTotal	Direction	AvgDirection	Where
Order	50%	0%	0%	50%	0%	50%
Trend	80%	60%	80%			60%
Location	100%	0%	0%	50%	50%	
Spread	67%		0%			
Outlier	50%	100%	50%	100%	100%	
Order:Location	50%	17%	50%	50%	33%	
Order:Spread	44%		33%			
Trend:Spread	50%		33%			

Table 5.4: Percentage of how many contrasts are in line with the hypotheses.

5.2. Distance

When annotators over- or underestimated their judgements, it means that the judgements have some distance to the ground truth. When the over- or underestimation is not corrected, and is a result of the queryset having certain conditions, it is seen as a context effect. To measure the distance to the ground truth, the ground truth is necessary. However, the problem is that the ground truth of the judgements is not known before the judgements are made, which makes it hard to see if the judgements are over- or underestimated. And if they are over- or underestimated, it has to be of such a magnitude that is bigger than the natural subjectivity of different annotators. For analysing the distance in this section, the golden judgements as described in 3.5.1 will be used as ground truth.

As discussed in the previous section, change can be used as a metric to indicate that a potential context effect occurs. Annotators changing their judgements is not necessarily a problem as long as the judgements are close to the ground truth. When annotators change their previous judgements, it means that the annotator is aware of the over- or underestimation of the previous judgements. The annotator corrects the context effects by changing the judgements. When context effects occur which are not corrected, thus not changed, it means that the context effects have an impact on the set of judgements. Therefore, the hypothesis is: When changes are made, the context effects are corrected, which results in judgements closer to the ground truth.

Not only change can be used as a metric, but also the factor levels themselves. After the judgements are made, the set of judgements belongs to certain conditions which are translated into the factor levels. If the distance to the ground truth is different among certain factor levels, it is another indication of a context effect.

5.2.1. Change

The hypotheses related to distance and change is that when annotators change previous judgements, the final judgements will be closer to the ground truth.

For analysing the distance with change as a metric, the distance to the ground truth is measured before and after changes are made. Each queryset consists out of 15 candidates and thus 15 Finescores, these Finescores are split into a set containing the Finescores before changes are made, and a set containing Finescores after changes are made. The latter one is the set that is submitted by the annotator. The first one is based on the logs. To indicate the distance to the ground truth, two metrics are used:

- **Pearson correlation:** For each queryset, the Finescores before and after change are separately used and compared to the ground truth. The Pearson correlation, as described in equation 3.1, is used to indicate the correlation with the ground truth.
- **Root-Mean-Square Error** The Root-Mean-Square Error (RMSE) is used to indicate the distance to the ground truth for each Finescore in the set. It is calculated with equation 3.2. The RMSE is calculated for both the set of Finescores before and after the change.

The comparison of two of the exact same querysets judged by the same annotator at another time span, is also called the intra-rater agreement. Flexer [18] studied the intra-rater agreement for a time

Pearson	MTurk
Mean	0.36
Median	0.46
Min	-0.61
Max	0.89
SD	0.36

(a) Pearson correlation inter-rater agreement MTurk

Pearson	After
Mean	0.89
Median	0.97
Min	0.10
Max	0.99
SD	0.18

(b) Pearson correlation before-after set

Table 5.5: The inter-rater agreement between annotators of the MTurk data compared with the intra-rater agreement for the before and after change set.

Pearson	Before	After
Mean	0.469	0.480
Median	0.514	0.519
Min	-0.582	-0.582
Max	0.927	0.927
SD	0.337	0.336

(a) Pearson correlation

RMSE	Before	After
Mean	27.06	27.23
Median	25.87	25.62
Min	13.52	13.52
Max	48.67	52.92
SD	8.22	8.73

(b) Root-Mean-Square Error

Table 5.6: Before is indicating the set of Finescores before the change, after is indicating the set of Finescores after the change. Both sets are compared to the ground truth and translated into the Pearson correlation and the Root-Mean-Square Error statistics.

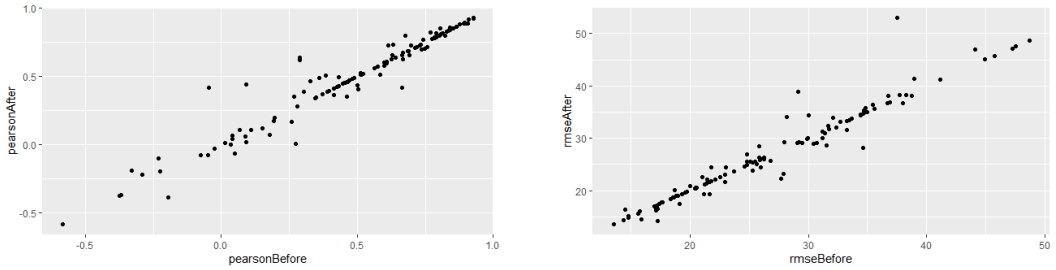
span of 2 weeks. This is clearly different than the timespan between the before and after set, which is a matter of minutes. However, the study showed that the intra-rater agreement correlation for a 2 weeks interval was higher than the inter-rater agreement, but not significantly higher. Due to the relatively short time period between the before and after set, it is expected the intra-rater agreement is higher than the inter-rater agreement. This is clearly the case when looking at the values in table 5.5.

The general statistics for the Pearson correlation to the ground truth can be found in table 5.6, the values are plotted in figure 5.1a. The clear line is due to the querysets not having any changes. The querysets which do have a clear difference between before and after the change, are querysets having more than 5 changes or a few changes which are large. For example, the four points above the line consist out of three querysets with more than 5 changes, and one queryset with two changes larger than 50. When comparing all correlation values to the inter-rater agreement correlation values, they are higher. The reason for this is probably due to the comparison to the ground truth instead of other annotators. The ground truth is averaged over all other annotators, meaning there are fewer outliers or extreme Fine scores which can lead to lower correlation. The general statistics show that the correlation to the ground truth is slightly better after the changes are made. When conducting a paired t-test, it confirms that the mean is smaller than 0. However, the differences are not significant so the null hypothesis can not be rejected, see table 5.7. A paired t-test is chosen because the comparison is made between the means of two related group of samples.

The same procedure is done for the RMSE values. The RMSE values are shown in figure 5.1b. The differences are not significant, see table 5.7, there is even less confidence than the Pearson correlation values to reject the null hypothesis. This is confirming the general statistics of table 5.6.

5.2.2. Factors

Not only the change, but the factors themselves can be used as a metric to see if over- or underestimation occurs in certain conditions. The factor levels representing the conditions are used as independent variables in the model. Distance is again measured with the Pearson correlation and the RMSE. In table 5.8, the estimated marginal means of the contrasts is shown. The model consists out of the main effects and the significant interaction effect according to an Anova. The contrasts which



(a) Pearson values

(b) RMSE values

Figure 5.1: The values of the Pearson correlation (a) and RMSE (b) between the before and after change sets.

	t	df	p-value	Mean	95% Conf interval
Pearson	-1.306	119	0.1939	-0.1078	[-0.0271, 0.0055]
RMSE	-0.827	119	0.4097	-0.1668	[-0.5661, 0.2324]

Table 5.7: Paired t-test results for the Pearson correlation and RMSE to the ground truth between the before and after change sets.

are around the significance level are in bold.

From the table, it can be seen that only two of the contrasts are around the significance level. The amount of data in these levels is quite low, with Outlier High having 31 querysets and Outlier Low 24 querysets, together with the interaction Trend:Outlier having 6 querysets in the Linear:High combination and 12 querysets in the Linear:Low combination. Taking in mind this low amount of data and considering the chance of Type I errors, the table does not show interesting results.

5.3. Summary

This chapter addresses context effects in two ways, whether they are recognizable and what the impact is. Change is used as a metric to see if annotators correct their judgements when they are aware of the over- or underrating of their previous judgements. How much the over- or underrating is, is based on the distance which is calculated with the Pearson correlation to the ground truth and the RMSE to the ground truth, both before and after change. For both the change and the distance, hypotheses are made which are checked whether they are in line with the data or not. To see if the data is in line, linear mixed-effects models are made with the factors as independent variables. From the models, the estimated marginal means are calculated for each factor level in the model. To see if the data is in line with the hypotheses, the contrasts between factor levels of the estimated marginal means are calculated and checked with the hypotheses. There is a distinction made between soft and hard evidence, the soft evidence is indicating the contrasts which are in line with the hypotheses, but not significantly. The hard evidence is showing the significant contrasts with a p-value below 0.05.

For both change and distance, there is no convincingly hard evidence. However, the soft evidence showed that some of the factor hypotheses are in line or with the data, or the data showed the opposite of the hypotheses. The Outlier factor seems to be in line with most of the hypotheses, while the Order factor shows the opposite. However, most of the factors seem to be in the middle of what was expected. The distance to the ground truth before and after change is indeed closer when changes are made, however, not significantly.

Factor	Contrast	Pearson		RMSE	
		Emmean	p-value	Emmean	p-value
Order	H2L - Random	-2,558	0,536	-0,04183	0,8784
	H2L - L2H	-2,458	0,4629	-0,03675	0,8701
	Random - L2H	0,101	0,999	0,00508	0,998
Trend	Exp - Linear	5,1357	0,5726	0,06431	0,9743
	Exp - Flat	3,3708	0,8288	0,11803	0,8425
	Exp - Log	3,2761	0,8676	0,12558	0,8495
	Linear - Flat	-1,765	0,889	0,05373	0,9304
	Linear - Log	-1,8596	0,9123	0,06128	0,931
	Flat - Log	-0,0946	1	0,00755	0,9998
Location	High - Middle	1,48	0,8601	0,0137	0,9922
	High - Low	-0,945	0,9377	0,1267	0,5397
	Middle - Low	-2,425	0,7419	0,1131	0,6258
Spread	High - Middle	-0,999	0,8884	-0,1036	0,4864
	High - Low	-3,36	0,2885	-0,0607	0,8335
	Middle - Low	-2,361	0,4788	0,0429	0,8684
Outlier	High - None	-3,77	0,1865	0,0724	0,6998
	High - Low	**-6,46**	**0,0565**	0,0888	0,7224
	None - Low	-2,69	0,4624	0,0163	0,9829
Trend:Location	Exp(High - None)	5,883	0,7897	-0,1582	0,8748
	Exp(High - Low)	10,344	0,2385	0,1452	0,8224
	Exp(None- Low)	4,461	0,8999	0,3034	0,7015
	Linear(High - None)	-8,535	0,1164	0,2647	0,2897
	Linear(High - Low)	-10,975	0,0522	**0,4232**	**0,0608**
	Linear(None- Low)	-2,44	0,7889	0,1584	0,5014
	Flat(High - None)	1,935	0,7888	0,1021	0,6714
	Flat(High - Low)	2,381	0,6212	-0,0337	0,9541
	Flat(None- Low)	0,446	0,9872	-0,1258	0,4826
	Log(High - None)	6,637	0,3393	-0,1539	0,6761
	Log(High - Low)	-5,531	0,662	-0,0277	0,9931
	Log(None- Low)	-12,168	0,116	0,1262	0,8382

Table 5.8: The estimate and p.value for the model with distance as response variable and the factors as IV.

6

Limitations

During the process of this work, some limitations are faced. Most of the limitations are dealt with, but the decisions made in the process have possibly influenced some of the results. With the knowledge today, a reflection can be made on the decision-making process during work.

The limitations worth mentioning are:

- **Factor distribution for reliability:** Before analysing data, the data needs to be reliable. In the first iteration of collecting judgements, only the extreme factor levels were chosen. Especially the Order factor could influence the reliability results, with in the first iteration only having the H2L and L2H order. The MIREX judgements are collected in random Order, which means that the reliability comparison in the first iteration was: MIREX random vs. MTurk H2L and L2H. Especially when taking in mind the research questions of this work, context effects in factor levels, the comparison is not reflecting the right question of reliability. To solve this issue, the random Order of the same queryset as H2L and L2H is also collected, with the result of having another factor distribution of the data than expected.

- **General factor distribution:** Not only the Order factor could have been distributed otherwise from the start, but also the general distribution for all factors. It was not really clear from the start how many querysets were needed and if they would show interesting results. Due to limited resources, only the extreme querysets were created at the first iteration, resulting in 40 querysets. Due to the reliability comparison for the Order factors, another 20 querysets in random order were added, leaving a distribution according to table 3.3. During the process, there were signs of results going into a certain direction, but there was not enough data. The decision was made to add another 60 querysets with the same factor distribution as the first 60 sets, meaning each combination has 2 different querysets. With the knowledge today, the distribution of factor levels could have been different from the start. For example, add a middle level to the Location or Spread factor. This would result in a more varied dataset, another decision can be to choose a less varied dataset with more annotators judging the same combination. For both decisions, there are arguments to support it, but it would be better to choose one of them. The dataset used in this work has a bit of both.

- **Subjectivity in context effects:** It is clear from literature that subjectivity is involved in music similarity judgements. However, the question arises of how much subjectivity is natural, and how much is influenced by context effects. And even if it is known, how much subjectivity is allowed? There is not a clear line in these kinds of questions, which makes the area a bit vague. Due to this subjectivity, the road to the research question is changed during the process. Starting with recognizing context effects by modeling the Fine scores themselves, which did not show significant results. Looking at the Fine scores itself changed to measurements which can indicate that a context effect occurs. By diving into the thinking process of an annotator, change

was chosen to indicate a potential context effect. The change ended up being a big part of the thesis, while it was initially not a big part of the plan.

- **Generated querysets for context effects:** The generated querysets are distributed according to the factor levels. While creating the querysets, the focus was on which candidates together are belonging to a factor level. When comparing factor levels for context effects, the best way is to compare the exact same candidates. This is only possible for the Order factor which has H2L, Random and L2H with exactly the same set of candidates. When comparing, for example, the Location factor with levels high and low, the comparison is made between two different sets of candidates which is not a solid comparison when taking in mind that it could be possible that the high median is 99, and the low median is 32. This is a comparison between an extreme high level and a high low level. It is of course not possible to create two querysets with exactly the same candidate belonging to both a high and low Location, but a better way would be to have a part of the queryset with exactly the same candidates. In this way, the comparison can be made between the part with the same candidates which are judges by the annotator in a set belonging to different factor levels. Another way would be to make more levels or not labeling querysets to levels, but looking at the value of the factor itself. However, for both ways a lot more data is needed.

- **Trend calculation:** The calculation of the Trend factor has been changed throughout the process. With the creation of the querysets, the querysets were ordered according to the H2L and L2H Order levels. When having a clear order, the trend can be calculated with the exponential and logarithmic equation. However, after the new judgements were collected, there was a lot of randomness in the set of candidates. In a random set of 15 candidates, the exponential and logarithmic equations have high errors when fitting the data. Therefore, the calculation of the Trend factor was changed to the equations from section 3.2.1, with the consequence of having two different calculation methods of the Trend factor at the creation of the querysets and at the analysis part. Although it did not seems to have a lot of influence on the results, it would have been better to use the last method from the start.

- **Confounding factor levels:** Some factor levels are confounding, which was expected at the creation of the querysets, but the decision was made to leave it as is. For example, an exponential Trend has most of the values at the low side of the queryset range. Combining this with a high Spread, there is a high chance that the queryset has a low Location. The confounding factor levels could have been filtered at the start, but this is not done because it excludes the extreme sets which are interested for the magnitude of effects. Creating querysets with a low chance of existence in a real scenario is useful to analyse the extremes, but a trade-off has to be made between the extremes and the usefulness in a real scenario. For this thesis, the extremes were chosen.

7

Conclusion

This chapter addresses the conclusions about the two following questions: 1) Are context effects measurable in the current annotation protocol of MIREX?, and 2) What is the impact of context effects in the current annotation protocol? After the conclusions, the future work is discussed.

7.1. Conclusion

To be able to answer both research questions, the data collected needs to be reliable. The MIREX judgements are collected by experts, or people who have affiliation with MIREX or the AMS task. The judgements in this work are collected with the crowdsourcing platform MTurk. Crowdsourcing is known to have a lower quality when not used carefully. To filter the low-quality judgements, quality control measures are taken. The MTurk judgements are collected with the same annotation protocol as the MIREX judgements. They are compared with four different measurements: 1) Pearson and Spearman correlation, 2) RMSE, 3) Pairwise Fine score, and 4) Factor comparison. The correlation and the pairwise comparisons seem to be highly comparable to the 2006 MIREX data. The correlation seems to be even better with the MTurk data, however, this could be due to extreme orders used in the MTurk data. The RMSE is better between annotators in the 2006 MIREX data, the reason could be, again, due to the extreme querysets used in the MTurk data. For the factor comparison between the MIREX and MTurk, there is no clear reference on how much they should differ, but the data seems to be in line with each other.

Overall, music similarity judgements collected with crowdsourcing seem to be reliable. When considering the subjective nature of similarity judgements, the differences between annotators obtained with crowdsourcing, and compared to the differences between annotators from MIREX seem to be negligible. However, this work does not have a lot of data. The low amount of data could have influenced the results, but the results given are promising considering cost and time.

After it is determined that the crowdsourced judgements are reliable, the research questions can be answered. For the first question, change is taken as a measurement. When annotators recognize that the judgements given, are over- or underrated, they will correct the judgements by changing them. Around half of the annotators in the MTurk data changed at least one of the judgements in the queryset. Extensive information about change in the MIREX data is unfortunately not available, which means there is no reference of what the standard is. Hypotheses are made for each of the factors Order, Trend, Location, Spread and Outlier, and for some factor interactions. To see whether the data is in line with the hypotheses, two kinds of evidence are presented, 1) soft evidence, and 2) hard evidence. Unfortunately, hard evidence is not really found which could have different reasons, the two main reasons are: 1) not enough data or 2) there is no difference in change among factors. However, the soft evidence shows some signs of a difference in change. The amount of changes seems to be in favour of the Trend and Outlier hypotheses, but not convincing for the rest of the factors. The Order factor and

the Average Total change response variable are even showing the opposite of the hypotheses. Overall, there is no hard evidence to say that change is different among factor levels.

The second question is about the impact of context effects. The impact is measured in distance to the ground truth which is given by the Pearson correlation and the RMSE. The change is used as a metric to recognize a potential context effect. The distance to the ground truth is analysed for the judgements before the change, and after the change. Both correlation and RMSE are showing closer Fine scores to the ground truth after a change, however, this difference is not significant. When using the factors as a metric for distance, there is also no significant distance among different factor levels. Overall, there is no hard evidence found that differences in distance exist among factor levels.

Both research questions are used to obtain a better understanding of occurring context effects in the current annotation protocol of the AMS task in MIREX. The goal is to give insight into how much of the disagreement among annotators is due to a natural subjectivity and how much is due to context effects in the annotation protocol. Unfortunately, no significant results are found, however, signs of context effects are measured and need to be further explored.

7.2. Future Work

This work has explored context effects in the current annotation protocol for the AMS task in MIREX. During the process and due to the limitations, additional questions are raised. This work was exploring the existence of context effects, but should be further studied to make significant conclusions. First of all, the presumption is made that, due to the limited amount of data, significant results were not obtained. At the start of the work, it was not known if the new collected data would show observations considering context effects. Therefore, three iterations are made to be able to make adjustments when the results were going into certain directions. Although some adjustments are made, it seems that still more data is needed to get significant results. The amount of data used in this work shows signs of context effects, for significant results, more data should be used.

The created instances are made based on balancing the factor levels, however, the distribution of the factors levels are not reflecting the MIREX distribution. A more extensive exploration of the factors would result in a better reflection of the MIREX data. If certain factors are not common in the MIREX data, it raises the question of how important it is and how much influence it has on the results. A more extensive exploration of the distribution of the MIREX factors should be done to obtain better insights of the importance of certain factors.

This work is only exploring whether context effects exist and if they have an impact on the judgements. In the end, it is about the system performance ranking of the AMS task. Even when context effects exist, it needs to be studied if they influence the system ranking. It could be possible that judgements are affected by context effects, but of such a small magnitude that system rankings are not affected.

The data obtained in this work is collected with crowdsourcing. Although the reliability of Fine scores in the current annotation protocol was not studied in current literature for the AMS task, it was not the main focus in this work. Studies have shown alternatives for crowdsourcing the AMS task, but the reliability of the Fine scores in the current annotation protocol seems comparable to the MIREX data. In this work, the reliability is not studied extensively with a lot of data, but the result seems promising. When a more extensively study about the reliability shows comparable results, MIREX could use crowdsourcing instead of experts for collecting similarity judgements.

Bibliography

[1] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81, 2013.

[2] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for trec relevance assessment. *Information processing & management*, 48(6):1053–1066, 2012.

[3] Omar Alonso, Daniel E Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. In *ACM SigIR Forum*, volume 42, pages 9–15. ACM New York, NY, USA, 2008.

[4] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800, 2011.

[5] Douglas G Bonett and Thomas A Wright. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1):23–28, 2000.

[6] Chris Buckley, Gerard Salton, and James Allan. The smart information retrieval project. In *Proceedings of the workshop on Human Language Technology*, pages 392–392, 1993.

[7] Cyril Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*. MCB UP Ltd, 1967.

[8] Cyril W Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1991.

[9] Philip B Crosby. *Quality is free: The art of making quality certain*. Signet Book, 1980.

[10] J Downie, Kris West, Andreas Ehmann, and Emmanuel Vincent. The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview. 2005.

[11] J Stephen Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.

[12] J Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[13] Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170, 2018.

[14] Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[15] Michael Eisenberg and Carol Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, 39(5):293–300, 1988.

[16] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[17] Arthur Flexer. On inter-rater agreement in audio music similarity. In *ISMIR*, pages 245–250. Citeseer, 2014.

[18] Arthur Flexer and Taric Lallai. Can we increase inter-and intra-rater agreement in modeling general music similarity?. In *ISMIR*, pages 494–500, 2019.

[19] Anatoliy A Gruzd, J Stephen Downie, M Cameron Jones, and Jin Ha Lee. Evalutron 6000: collecting music relevance judgments. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 507–507, 2007.

[20] Derek L Hansen, Patrick J Schone, Douglas Corey, Matthew Reid, and Jake Gehring. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 649–660, 2013.

[21] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[22] Mu-hsuan Huang and Hui-yu Wang. The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the American Society for Information Science and Technology*, 55(11):970–979, 2004.

[23] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67, 2010.

[24] Karen Sparck Jones and Peter Willett. *Readings in information retrieval*. Morgan Kaufmann, 1997.

[25] M Cameron Jones, J Stephen Downie, and Andreas F Ehmann. Human similarity judgments: Implications for the design of formal evaluations. In *ISMIR*, pages 539–542, 2007.

[26] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[27] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[28] Jin Ha Lee. Crowdsourcing music similarity judgments using mechanical turk. In *ISMIR*, pages 183–188, 2010.

[29] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 129–138, 2012.

[30] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[31] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan Schwamb, Chris Lintott, and Arfon Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, 2013.

[32] Afra J Mashhadi and Licia Capra. Quality control for real-time ubiquitous crowdsourcing. In *Proceedings of the 2nd international workshop on Ubiquitous crowdsouring*, pages 5–8, 2011.

[33] Thomas Mussweiler and Fritz Strack. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2):136–164, 1999.

[34] Gabriel Parent and Maxine Eskenazi. Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[35] Lorraine M Purgailis Parker and Robert E Johnson. Does order of presentation affect users' judgment of documents? *Journal of the American Society for Information Science*, 41(7):493–494, 1990.

[36] G Salton. The smart system. *Retrieval Results and Future Plans*, 1971.

[37] Parnia Samimi and Sri Devi Ravana. Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: a review. *The Scientific World Journal*, 2014, 2014.

[38] Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[39] Eric Schenk and Claude Guittard. Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics Management*, (1):93–107, 2011.

[40] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

[41] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

[42] Fritz Strack and Thomas Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, 73(3):437, 1997.

[43] Amos Tversky. Gati,(1978),'studies of similarity'. *Cognition and categorization*, pages 79–98.

[44] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[45] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[46] Julián Urbano, Jorge Morato, Mónica Marrero, and Diego Martín. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *ACM SIGIR workshop on crowdsourcing for search evaluation*, pages 9–16. ACM New York, 2010.

[47] Julián Urbano, Markus Schedl, and Xavier Serra. Evaluation in music information retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369, 2013.

[48] Ellen M Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*, pages 355–370. Springer, 2001.

[49] Kerri Wazny. "crowdsourcing" ten years in: A review. *Journal of global health*, 7(2), 2017.

[50] Daniel Wolff. *Spot the Odd Song Out: Similarity Model Adaptation and Analysis using Relative Human Ratings*. PhD thesis, City University London, 2014.

[51] Yunjie Xu and Dong Wang. Order effect in relevance judgment. *Journal of the American Society for Information Science and Technology*, 59(8):1264–1275, 2008.

[52] Dongqing Zhu and Ben Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 17–20, 2010.