Perceptions of Diversity in Electronic Music: the Impact of Listener, Artist, and Track Characteristics

LORENZO PORCARO, Music Technology Group, Universitat Pompeu Fabra, Spain

EMILIA GÓMEZ, Music Technology Group, Universitat Pompeu Fabra, Spain and Joint Research Centre, European Commission, Spain

CARLOS CASTILLO, Web Science and Social Computing Group, Universitat Pompeu Fabra, Spain

Shared practices to assess the diversity of retrieval system results are still debated in the Information Retrieval community, partly because of the challenges of determining what diversity means in specific scenarios, and of understanding how diversity is perceived by end-users. The field of Music Information Retrieval is not exempt from this issue. Even if fields such as Musicology or Sociology of Music have a long tradition in questioning the representation and the impact of diversity in cultural environments, such knowledge has not been yet embedded into the design and development of music technologies. In this paper, focusing on electronic music, we investigate the characteristics of listeners, artists, and tracks that are influential in the perception of diversity. Specifically, we center our attention on 1) understanding the relationship between perceived diversity and computational methods to measure diversity, and 2) analyzing how listeners' domain knowledge and familiarity influence such perceived diversity. To accomplish this, we design a user-study wherein listeners are asked to compare pairs of lists of tracks and artists, and to select the most diverse list from each pair. We compare participants' ratings with results obtained through computational models built using audio tracks' features and artist attributes. We find that such models are generally aligned with participants' choices when most of them agree that one list is more diverse than the other. In addition, we observe how differences in domain knowledge, familiarity, and demographics influence the level of agreement among listeners, and between listeners and computational diversity metrics.

CCS Concepts: • **Information systems** → **Information retrieval**; • **Human-centered computing** → User studies.

Additional Key Words and Phrases: music information retrieval; diversity; reliability;

1 INTRODUCTION

Research on diversity in Information Retrieval (IR) typically seeks diversity as a multiple criteria objective, or criteria based on multiple diversity axioms. A diverse set of items should be, among other things, a list in which each item is *novel* with respect to those previously seen, in which items cover different potential *intents* of the user, and convey different *aspects* of the information sought [3]. Computational approaches, adopted in a number of IR contexts, enable the estimation of measurements that can serve as proxies of these criteria within specific tasks, such as subset selection [19, 51], ranking [14, 32], and recommendation [43]. A complementary approach to diversity seeks to understand and ideally model what users *perceive* as diversity, a challenge that is addressed through user studies and fraught with difficulties [62]. Several works confirm that characteristics such as demographics [58], personality traits [13], or domain knowledge and expertise [41] influence individuals' diversity needs and perceptions. In general, what a system considers to be *diverse* may be perceived differently by people interacting with it.

The present study investigates the interrelationship between diversity aspects related to listeners, artists, and tracks in the music domain. These actors have been represented as the threefold nature of the music phenomenon [52]: "*as an arbitrarily isolated object* [a track], *as something produced* [by an artist], *and as something perceived* [by a listener]". Our research aims to deepen our understanding of how different diversity facets interplay in the music field and how are they experienced: "[..] *subjectivity as coloured by embodied experience, which in human societies is mediated*

by such social differences as gender, class, race and ethnicity – so that subjectivities, experience and embodied experience are not everywhere the same."[8].

We designed a user study to investigate diversity in the music domain, specifically in the Electronic Music (EM) genre. The study was divided into three parts. In the first one, the stimulus consisted of two sets of four tracks each. We asked participants to briefly listen to the tracks and evaluate which, among the two sets, they perceived to be more diverse. In the second part, the stimulus was two sets of photos of four artists each, along with some basic information about them. We asked participants to indicate which set was more diverse according to socially salient attributes (e.g., age and gender). In the following, we refer to part one as the *TRACK* task and part two as the *ARTIST* task. A third part which we call *COMBO* combined the *TRACK* and *ARTIST* tasks.

We built a computational model to automatically estimate diversity by considering distance measures between tracks based on specific music-related features extracted from the audio signal, as well as distance measures between artists based on their attributes. Then, we compared participants' ratings to this model, performing inter-rater reliability analysis (IRR) and metric reliability analysis (MR). Additionally, we compared the results from all participants against subgroups created by clustering participants based on their EM domain knowledge, familiarity, and/or demographics. Figure 1 depicts the high-level structure of the experiment.
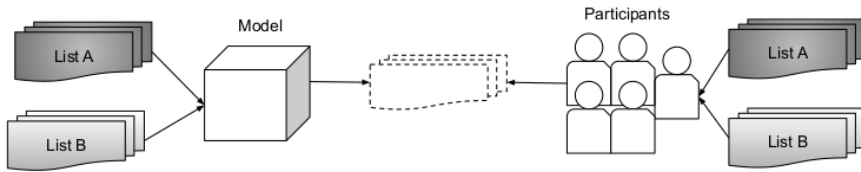

Fig. 1. High-level view of the study. On the left side, two lists are provided to a computational model designed to identify the most diverse one, while on the right side the same lists are provided to a group of people. By analyzing the agreement among the two processes, we aim at highlighting the key factors in diversity estimations.

The purpose of this analysis is to answer the following research questions:

(1) **RQ1. To what extent tracks' audio features and artists' attributes can be used to assess perceived diversity?**
(2) **RQ2. To what extent domain knowledge and familiarity influence participants' perceptions of diversity?**

Our main finding regarding RQ1 is that computational methods can to some extent be used to compare diversity between two lists, especially when there is a strong agreement among listeners that one list is more diverse than the other. Our main findings regarding RQ2 are that domain knowledge and familiarity with a music genre are partially aligned, and that in general listeners with more domain knowledge agree less with automatically extracted metrics, and disagree more with each other.

We develop these and other findings in the rest of the paper as follows: Section 2 provides an overview of how diversity has been approached from an information processing perspective, and surveys diversity and agreement studies performed in the Music Information Retrieval field. Following, Section 3 reports the method used in this work, describing the user study design, while Section 4 the considered features and metrics. Section 5 presents the results of our analysis, discussed together with their limitations in Section 6. Conclusions are presented in Section 7.

2 BACKGROUND AND RELATED WORK

Next, we introduce the concept of diversity for then reviewing its related literature in ranking and recommendation. Afterwards, focusing on the Music Information Retrieval (MIR) literature we first present diversity-related user studies on music listening, and we then review the work done in the analysis of annotators' agreement.

2.1 The Value of Diversity

The definition and operationalization of diversity have been the object of study of scholars from several disciplines through history, ranging from ecology to information science, economic and cultural studies [69]. Such a plethora of approaches investigated several concepts of what can be intended as *diversity*, which according to the context can be defined and measured differently, with different implications as discussed in detail by Steel et al. in [68]. Nonetheless, what generally is shared among academic communities, but also public and private institutions, is the importance of fostering diversity as a tool to enhance creativity but also productivity [69], to promote pluralism and equality [70], to make people aware of different viewpoints and facilitate the public debate [36].

Even if these benefits have been proven in several areas, simultaneously major concerns have emerged due to the lack of diversity, and Artificial Intelligence (AI) is a notable example of a field in which we are witnessing a diversity crisis [31, 76]. Indeed, AI systems — an umbrella term wherein also Recommender Systems can be included [40] — have already been proven to reinforce hegemonic biases and discrimination in different applications, in fields such as Computer Vision [60] and Natural Language Processing [5]. To contrast these harmful effects, in the process of design, development, implementation, and evaluation of these systems, diversity has been already identified as a key requirement to build trustworthy AI systems [54].

In the music field, scholars from music psychology, sociology, and ethnomusicology questioned under different perspectives the value of diversity [15, 34, 39], stressing the importance of pushing the boundaries of researching and learning beyond Western music traditions. On the contrary, few examples in the MIR literature can be found where multicultural analysis has been carried out, e.g. [66], and the lack of diversity in this field has been already acknowledged in different contexts [8]. However, recently the effort of the MIR community in promoting initiatives like Women in MIR[1], and also organizing within the International Society for Music Information Retrieval (ISMIR) conference, the main forum of MIR practitioners, a special call for Cultural Diversity papers[2], make evident how nowadays diversity has been acknowledged as core topic on which more research and awareness is needed.

2.2 Diversity in Ranking and Recommendation

Information Retrieval systems have benefited from the introduction of evaluation practices in which diversity is key to determine whether results are useful for end-users. In ranking scenarios, the Probability Ranking Principle (PRP), affirming that documents should be ranked by their probability of being relevant in decreasing order, has been shown to be unable to deal with concepts such as ambiguity and redundancy [12]. In the well-known example in the IR community of the ambiguous query "Jaguar", a retrieval system for satisfying the possible information needs of a user is expected to return documents related to, e.g., an animal, a carmaker, and a brand of guitars, and diversity may aid in expanding the PRP offering a new perspective from which to rank relevant documents [14]. Diversification techniques recognize that the relevance of a document depends not only on its

---

[1]https://wimir.wordpress.com
[2]https://ismir2021.ismir.net/cfp

nature but also on its relationships with other documents [2]. These techniques are often based on axiomatic approaches describing the properties that diversification systems and diversity metrics should satisfy [3, 32].

Parallelisms between ad-hoc IR and Recommender System (RS) diversification techniques have naturally emerged, given that the recommendation task can be seen as an IR task [73]. Similarly to what happens in ranking settings, redundancy in RS can create a so-called *portfolio effect*, i.e., recommendations provided to users are too similar to each other [10]. As a consequence, RS research has highlighted the relevance of going beyond accuracy metrics when evaluating the retrieved results to increase the overall users' satisfaction [50, 79].

Centering on user studies, Sanderson et al. [63] compared different retrieval systems and found that users tend to prefer systems that, according to the considered diversity metrics, return more diverse outcomes. A similar result is presented in [11], where authors showed that users are more satisfied with recommendations provided by the system optimized to maximise the proposed diversity metric. Such results are confirmed in [77], which proves that diversification techniques help in reducing users' choice difficulty and choice overload. Moreover, in [21, 78] authors found that users' satisfaction is positively correlated with the diversity of the provided recommendations.

Nonetheless, even if in the last decades the IR and RS communities have actively researched diversity, there is still an ongoing debate around whether such diversity metrics are effectively beneficial from a user point of view [62]. Generally, offline evaluation can be easily performed to understand points of strength and weakness of retrieval systems, while user studies are more difficult to execute and rarer in the literature [47, 59]. The interactions between humans and retrieval systems are particularly challenging to analyze due to the differences among users' characteristics and needs, and to collect such data is a time-consuming task and expensive in terms of resources.

2.3 Diversity in Music Listening User Studies

Diversity has been empirically investigated by the MIR community in relationship with listening behaviours, mostly focusing on its application in recommendation scenarios. In [26], Ferwerda et al. grouped listeners by their country of origin, and then analyzed how different diversity needs are linked to each country's cultural dimensions [38]. Farrahi et al. [23] showed how recommender systems performance increases for users with greater diversity in terms of listened music genres. Furthermore, analyzing the diversity perceived by listeners when exposed to a recommendation list, a positive correlation has been found between the perceived diversity, recommendation attractiveness, and the possibility to discover new music [24]. The same authors also explored the link between musical sophistication and diversity needs, showing how a higher sophistication corresponds to more diversified listening behaviours [25]. In addition, through semi-structured interviews, authors in [61] examined what listeners look for when faced with diversity-aware music recommendations, characterizing participants' inner and outer diversity as diversity within and outside their own music preferences. In most of these studies, users' listening behaviours are characterized by their diversity, computed using as a proxy the number of different artists, tracks, or music genres listened to by users. In contrast with previous work, we focus on the exploration of what are the characteristics that are associated with perceived diversity in the music domain.

2.4 Human Annotations Agreement in Music Information Retrieval

Our study features heavily human annotations over music, which is why we present here some background on this topic, connecting to previous works from the MIR literature.

The analysis of the agreement between human annotators is considered, among other factors, to influence the reliability of scientific experiments, giving a more robust interpretation of the obtained results [72]. While in the IR field evaluation practices have a long tradition, and are

a subfield of research on their own, they are comparatively not as well established in MIR and most of the work done has focused on the concept of music similarity, more than diversity [65]. Indeed, because of its relevance in several MIR-related tasks, such as music classification, cover detection, and music recommendation, the MIR community has posed great attention in developing proper evaluation practices for music similarity [71]. Flexer et al. [28, 29] showed that due to the highly subjective, context-dependent, and multi-dimensional nature of music similarity, the lack of inter-rater agreement between annotators provides an upper bound of the performance of retrieval systems based on such notion. A recent study showed how selecting a controlled group of subjects and music material can lead to an increase in inter-rater agreement, but still questioning the validity of experiments on general music similarity [30].

From another perspective, the inter-rater agreement has been analyzed in the context of Music Emotion Recognition (MER), another task in which the high subjectivity in estimating emotion perceived when listening to music has limited the applicability of algorithmic solutions. Schedl et al. [64] analyzed the relationship between listeners' characteristics and emotion perceived while listening to classical music, founding a greater agreement for non-expert classical music listeners. Similarly, in [33] authors investigated the role of personal characteristics when classifying annotations related to music, focusing on the rock and pop music genres, showing that annotators have a higher agreement for basic emotions than for complex ones. A further example of annotators' agreement analysis is presented in [55], where agreement is used to evaluate the performance of information extraction methods for building a knowledge base in the music domain.

Following previous research, this study analyzes annotators' agreement to understand how diversity is perceived among different groups of listeners, studying also the agreement between them and diversity metrics as discussed in the following sections. On one hand, measuring metric reliability we seek agreement as confirmation of the proposed diversity metrics. On the other hand, using inter-rater reliability we evaluate if data are interpreted in the same way among the participants. This is in line with norms and guidelines for CSCW and HCI practice [48].

3 STUDY DESIGN

We designed an online survey for our user study. A facsimile of the survey, along with anonymized responses, is available with our data release.[3] The study was an activity of the TROMPA (Towards Richer Online Music Public-domain Archives) project[4] and followed the personal data management and ethical protocols approved within this project.

3.1 Recruitment and Informed Consent

We recruited participants by sharing a landing page from where to access the survey, using the authors' personal and institutional social networks (Twitter, Facebook, Reddit, and Instagram), and MIR-related mailing lists. Responses were collected from July through October 2020.

Participants were shown details of the survey and an information sheet describing the research objectives, methodology, risks, and benefits. They were informed that participation was voluntary and could be withdrawn at any point, and of their rights including the right to access, rectify, and delete their information. Personal data collection was limited to what the participant chose to disclose, as all demographic and background questions were labeled as optional. At the end of the survey, participants could opt-in to be asked further questions, and/or to receive the results of their study by e-mail; in that case, they could provide their email address. This part was also clearly marked as optional.

---

[3]https://zenodo.org/record/4436649
[4]https://trompamusic.eu

3.2 Survey Sections

The survey was composed of five main sections:

(1) *Participants' information collection.* Optional demographic information, with each question prominently labeled as optional: age, gender, skin tone according to the Fitzpatrick scale [27], continent origin, level of education. With regards to the choice of including skin tone, we agree with the work by Mitchell et al. [51] that while introducing a diversity metric it is critical to consider attributes "[...]*defined in light of human attributes involved in social power differentials, such as gender, race, color, or creed*". Under this lens, we consider skin tone as one of the social salient characteristics which may impact the perception of diversity. Nonetheless, we acknowledge that Fitzpatrick scale is an approach to skin classification that is not exempt from racial limitations, as discussed in [74]. As detailed in Section 5.1, the demographic information has been collected in order to 1) describe the population of the experiment, and 2) cluster participants in subgroups based on demographic characteristics.

(2) *Domain knowledge.* Optional information about participant's music background and familiarity with electronic music: whether they had any formal music training (yes/no), music playing engagement, taste variety, EM listening frequency, EM taste variety (5-Point Likert scale). This was accompanied by a test for estimating participants' familiarity with EM artists (detailed in Section 3.3).

(3) *Track Variety (TRACK) task.* Participants were asked to listen to two sets of audio tracks and to choose the most diverse in terms of music features. Tracks were shown by means of YouTube embeddings, without specifying additional information such as tracks' title or artist (see Figure 2). Participants were explicitly asked to not consider the tracks' popularity, their preferences, or their familiarity with the tracks, but to focus only on selecting the most diverse list within each pair. They were also not asked to listen entirely to every track, but to navigate through them as they normally would do when listening to music in a digital environment.

(4) *Artist Diversity (ARTIST) task.* Two lists of artists were shown to participants, then asked to choose the most diverse one in terms of socially relevant attributes. Artists were presented using close-up or half-length photos, avoiding when possible images where they are performing or acting (see Figure 2). Together with the photo, the name of the artist and the birthplace were shown. Again, participants were explicitly asked to not consider artists' popularity, their preferences, or their familiarity with the artists when selecting the most diverse list in each pair.

(5) *Combined (COMBO) task.* Two lists of tracks and corresponding artists were shown to participants, asked to indicate the set most diverse in terms of tracks' music features and artists' attributes at the same time. The information displayed for every list was the sum of the previous tasks. The indications for selecting the list given to the participants were the same: to not consider popularity, personal preferences, or familiarity when choosing the most diverse list.

While the *TRACK* task is the most realistic of all and corresponds more specifically to how people consume music, *ARTIST* and *COMBO* tasks instead can be considered experimental artifacts to complement the observations from the first. Several works in multimodal Music RS research, in addition to audio signals, exploit the so-called contextual information such as album covers [45], or artist biographies [56]. This is motivated by the fact that the content itself (i.e. the audio in the music domain) does not carry all the music information, and therefore methods including other types of data have recently emerged, also thanks to the advancement of Deep Learning techniques.

Similarly, we believe that to assess the diversity of a music list, the audio may be the component with the most influence on the perception but not the only one.

In each task (*TRACK*, *ARTIST*, and *COMBO*), participants had to compare four pairs of lists, each list formed by four elements (i.e., tracks, artists, and tracks and artists together). For each comparison, they had to select between the first list ("*List A*"), the second one ("*List B*"), or neither ("*I don't know*"). At the end of each task, participants were asked to provide feedback about their selection strategy, indicating the factors that most influenced their choices. To avoid order effect bias, we created four versions of the survey, switching the order of the *TRACK* and *ARTIST* tasks (parts 3 and 4 above), and the order of the lists within each pair. The survey was implemented using Google Forms and available in English, Spanish, and Italian: participants were able to select the preferred language for completing the survey on the landing page.
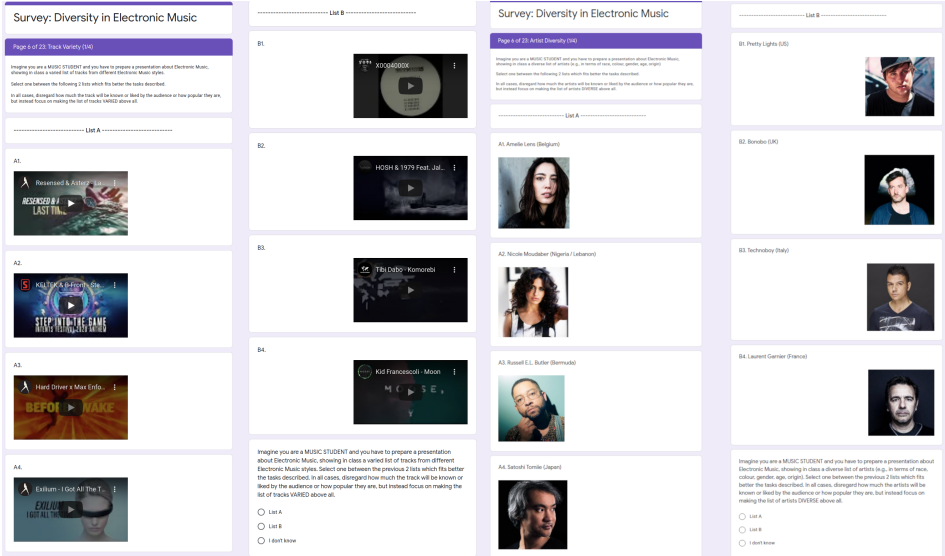

Fig. 2. Screenshot of the interface for the *TRACK* (left) and *ARTIST* (right) tasks.

3.3 Domain Knowledge and Familiarity

We have been able to estimate participants' familiarity with EM artists, independently from the previously self-declared answers, by designing a quiz presented at the end of the first part of the survey, in which they were asked to select which among a list of *n=30* selected artists they know, they do not, or they maybe know. The list was composed of 15 artists picked within the EM highlights of AllMusic,[5] a selection curated by music experts and including EM most representative artists, and 15 artists selected within the ones shown in the survey tasks, mostly less-known and not mainstream. Artists were presented in a randomized fashion to avoid order effect bias. As a result, we assigned to every participant a familiarity score, computed as an inverse-popularity weighted sum; this reflects the intuition that knowing several of the less popular artists implies more familiarity with EM.

---

[5]https://www.allmusic.com/genre/electronic-ma0000002572/artists

Specifically, the familiarity score was computed with the following procedure. First, we gather several social signals for every artist on the list: Spotify, Twitter, Facebook, Deezer, and Sound-Cloud followers, together with Spotify monthly listeners, and Facebook likes. We aggregate such popularity-like data using the geometric aggregation of popularity (GAP0) metric presented in [42], a non-linear method for popularity metric aggregation based on geometrical shapes. The list of artists and the corresponding GAP0 score is presented in Figure 3 (see Appendix A). Afterward, we rank artists by their score, and we assign an alternative score inversely proportional to their GAP0 popularity, ranging from $1/n$ for the most popular artist to 1 for the less popular. The final familiarity score assigned to every participant is a weighted sum of the artists' alternative score, where participants got: a) the entire score if the artist is known b) half of the score if the artist is maybe known, and c) zero if the artist is unknown. Finally, participants' familiarity score is normalized between 0 and 1.

3.4 Task Selection

Tracks and artists data have been manually selected by searching in knowledge bases such as *Wikipedia*, *AllMusic*, *MusicBrainz*, *Discogs*, and *Resident Advisor*, but also from artists' — websites, social media, and interviews in music magazines. The complete list of tracks and artists used for the *TRACK*, *ARTIST*, and *COMBO* tasks are reported in Appendix A (respectively Table 4, Table 5, and Table 6). The displayed items have been chosen to guarantee that the differences between lists were more pronounced in the former two out of four pairs, while more subtle in the latter two pairs. In the case of the *TRACK* task, such differences were based on tracks' audio and semantic features (see Section 4.1), e.g. showing a list where tracks have different tempo or are associated with various music genres, whilst in the case of the *ARTIST* task, diverse lists are where artists belong to different social groups. This strategy has been chosen for helping the participants become familiar with the task. Similarly, the *COMBO* task has been shown after the previous two tasks because of its higher complexity, since the comparison had to be based on both tracks' and artists' information.

Only solo artists have been included in the experiment. Indeed, even if they do not represent the totality of the EM scene, we avoid including duos or groups to facilitate the comparison between lists. Regarding the content included in the survey, while some of the tracks and artists are shared among different tasks, they have been selected independently in each case for being able later to analyze and discuss results separately, as presented in Section 5. As consequence, the lack of a major overlap of the lists among tasks made it not possible to analyze robustly how consistent were the participants' judgments across the three tasks, a factor excluded in the initial design of the experiment. The evaluation of the diversity using of the distance metrics, later introduced in Section 4.3, confirms our design strategy as discussed in Section 5.2.

4 FEATURES AND MEASUREMENTS

This section describes the metrics that we computed to estimate the diversity of lists of tracks and artists.

4.1 Track Features

Two groups of features for computing distances between tracks have been selected from standard features proposed in the MIR literature [65]: low-level, rhythmic, and tonal features (LL), and high-level features (HL). LL features are directly related to audio signal properties, while HL features represent aspects that are more understandable by human perception. Examples of features included in the LL group are BPM (Beats Per Minute), related to the tempo of the track, or HPCP (Harmonic Pitch Class Profile), describing the pitch and tonal content of the piece. In the set of HL features, we

consider mood-related features, such as "happy/sad" or "aggressive/relaxed", features describing the presence of voice or instrumental parts, or the darkness/brightness of a track. The complete list of features is included in Appendix A (see Table 7). Specifically, we used *Essentia* [6], an open-source library for audio and music analysis to extract the features from the audio signals. A detailed explanation of the features and of the algorithms used to extract them is publicly available.[6] Even if LL and HL features represent different characteristics of an audio track, there are some correlations among them. For instance, in the set of tracks used in the *TRACK* task, the HL features *danceability* and *party* are positively correlated with BPM (Pearson's $\rho = .41$, $p < .05$ and $\rho = .50$, $p < .005$), while *relax* is negatively correlated with it ($\rho = -.28$, $p = .12$). Similarly, dissonance and spectral entropy are positively correlated with *aggressivity* ($\rho = .38$, $p < .05$ and $\rho = .37$, $p < .05$).

4.2 Artist Attributes

We focused on four attributes for estimating the diversity of an artists' set: gender, origin, skin tone, and debut year. Gender and skin tone are two salient characteristics included in the image set perception study proposed in [51], and artists' birthplace was explicitly presented to the participants during the survey (see Section 3). Besides, we included the debut year as a characteristic representing part of the cultural aspects of the artists' music background. In the case of gender, we grouped artists into two groups: male and not-male, including female, non-binary, transgender, and any other gender different than male. The reason for this grouping was to compare a privileged gender within the music industry (men) against the other genders [22]. For the origin, we used as a proxy the artists' birthplace and then built three classifications: at a country-level, continent-level, and using regional groupings proposed by the United Nations in the context of the Millenium Development Goals.[7] For the skin tone, we first classified artists using the Fitzpatrick Scale, and then grouped them in white (skin type I and II) and not-white (skin type III, IV, V, and VI). Lastly, using the debut year, which represents a music aspect (a moment in electronic music history) but also a generational difference among artists, we divided artists into four groups: 1980s, 1990s, 2000s, 2010s. For every attribute, artist groups were balanced considering socio-political power differentials: male versus not-male, white versus non-white, developed regions versus developing regions, and recent versus non-recent artists.

4.3 Distance Metrics

The measurement of track and artist diversity was accomplished using two distinct metrics, respectively the cosine distance [67], and the Goodall distance [7]. This choice is motivated by the difference between the considered tracks' features and artists' attributes. Indeed, in the former case we estimated the distance using as input a vector composed of normalized numerical values, while in the latter we used categorical data. Cosine distance has the advantage of being 0 for identical vectors and 1 for orthogonal ones. The Goodall metric has the advantage of assigning a lower distance to a match when the attribute value is rare. For instance, in a set formed by ten artists, seven from Northern America and three from Latin America and the Caribbean, all other things equal, two artists coming from Northern America will score a higher distance in comparison to two artists coming from Latin America and the Caribbean, because of the larger presence in the set of the former in comparison to the latter.

The assessment of the most diverse list among each pair was accomplished using a two-step procedure:

---

[6]https://essentia.upf.edu/streaming_extractor_music.html
[7]http://mdgs.un.org/unsd/mdg/Host.aspx?Content=Data/RegionalGroupings.htm

(1) the pairwise distance between items in a list was computed (d_i), using the cosine metric for the tracks' features and the Goodall metric for the artists' attribute.
(2) using two aggregation operations, namely the minimum pairwise distance between items, $\min(d_i)$, and the average pairwise distance between items, $\text{avg}(d_i)$, we identified the most diverse list as the one with a higher value obtained after aggregating.

Such strategies were chosen following the work by Mitchell et al. on diversity and inclusion metrics for subset selection [51]. Using the $\min(d_i)$ mechanism, the most diverse list is assured to have all the items more diverse among each other than the other list, while in the case of $\text{avg}(d_i)$ the list with a higher diversity has on average items more diverse between each other. As observed in [51], these two mechanisms can be mapped to two approaches from social choice theory, namely *egalitarian* [4] and *utilitarian* [18].

5 RESULTS

This section presents the results of our survey, which are summarized and discussed in Section 6.

5.1 Participant's Statistics

After filtering out participants less than 18 years old, and participants who did not express any preference in any task (i.e., selecting always "*I don't know*"), we obtained a total of 115 answers.

Most of the participants are aged between 18 and 35 (61%), male (73%), come from Europe or North America (73%), have a bachelor's degree or higher (85%), and identified themselves to have a skin tone type I (pale white skin) or II (white skin) according to the Fitzpatrick scale (60%). In the light of these numbers, we can affirm that the population of our study is formed for the most part by subjects coming from WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies [37]. Regarding their music background and expertise, more than half of the participants indicated that they have received formal music training beyond the usual music lessons in school (58%). Moreover, using a 5-point Likert scale they declared on average to moderately engage in playing, DJ-ing, or producing any kind of music (M=3, IQR=4), to have high to moderately varied listening habits (M=4, IQR=1), to be EM listeners (M=5, IQR=1), and within EM to listen to different styles (M=4, IQR=2).

The familiarity test with EM artists (see Section 3.3) gave us another perspective to analyze participants. On average, they obtained a score of 0.3 (on a scale from 0.0 to 1.0) with a standard deviation of 0.2. Looking at the score distribution grouping participants by their self-declared frequency of EM listening, we notice that the more they listen to EM, the higher familiarity score they obtained. In fact, participants who self-declared to listen to EM with a high frequency (4 and 5 in the 5-point Likert scale) on average obtained 0.34 on the familiarity score. Ones with medium frequency got on average 0.19, while participants with low frequency 0.1, and the ones who almost no listen to EM obtained 0.05 (respectively self-declaring 3, 2, and 1 on the Likert scale). Besides, we analyze the correlation between the familiarity score and the other self-declared music information finding a positive correlation with the variety of EM listened (Pearson's $\rho = .40$, $p < .001$), the frequency of EM listening ($\rho = .39$, $p < .001$), the variety of music taste ($\rho = .22$, $p < .05$), and the frequency of music playing ($\rho = .19$, $p < .05$). On the contrary, we found no evidence of a correlation between receiving formal music training and the familiarity score ($\rho = -.08$, $p = .42$). These findings suggest that the familiarity score can be a further dimension to add when assessing participants' domain knowledge within EM, and it has the advantage of not being a self-declared attribute.

5.1.1 Participant Clustering and Demographic Grouping. We performed an automatic clustering procedure to group participants based on their EM domain knowledge and familiarity. To do that,

we used the self-declared information about the frequency of listening to electronic music and taste variety, and the familiarity score. We selected the k-medoids clustering technique (with k=3) [57], implemented in the *scikit_extra* Python package,[8] because of its lower sensibility to outliers in comparison to other techniques such as K-means. Additionally, we removed outliers identified as the points with a distance from the medoids greater than the average distance plus the standard deviation. In this manner, we have been able to define three groups of participants respectively with low (group G1, size N=39), medium (G2, N=27), and high (G3, N=30) domain knowledge. We additionally created four other groups based on demographic information for analyzing participants' ratings. First, we include in the WEIRD group (G4, N=84) participants coming from Europe and North America, holding a bachelor's degree or a higher title. The remaining participants were grouped in the non-WEIRD group (G5, n=24). Then, we created two other groups dividing participants by age: age between 18 and 35 (G6, N=70), and age over 35 (G7, N=42). While participants can belong to only one of the groups based on the domain knowledge (G1, G2, or G3), in the case of the demographic groups participants may have been included in either G4 or G5, and at the same time in either G6 or G7. This double-assignment is motivated by the purpose of analyzing different populations, divided by origin in group G4 and G5, and age in group G6 and G7, in which however a participant can coexist. Participants who chose to not disclose such demographic information at the beginning of the survey were not included in these groupings.

5.2 Agreement with Diversity Metrics

The results of the diversity estimated by means of the distance metrics are shown in Table 1, together with the participants' percentage agreement. In addition, Table 8 (see Appendix A) provides the results of the Mann-Whitney U test [46], to assess if differences between distances in the lists are statistically significant.

Regarding the agreement, we observe that in the *TRACK* task the greater agreement is obtained where the two aggregation mechanisms ($\min(d_i)$ and $\text{avg}(d_i)$, see Section 4.3) agreed on which list was the more diverse. Such finding is also confirmed by the statistically significant difference in the two first pairs of lists (*TRACK1* and *TRACK2*). In contrast, looking at the other two pairs (*TRACK3* and *TRACK4*) we observe a decrease in the agreement in participants' ratings where the aggregation mechanisms return contrasting results. In addition, for such pairs the result of the Mann-Whitney U test does not suggest a significant difference between the lists presented either in terms of LL or HL features. When asked for feedback on the selection strategy for this task, participants confirm that the following factors had a strong-medium influence on their choices: instrument and samples (83% participants), sub-genre or sub-style (82%), tempo/BPM (80%), feeling, emotion and mood (70%).

For the *ARTIST* tasks, we observe an analogous behaviour. Indeed, the greater agreement between participants is found for the first two pairs where there is a large difference in the Goodall metric (*ARTIST1* and *ARTIST2*), while in the other two (*ARTIST3* and *ARTIST4*), the agreement decreases. According to the participants' feedback on their selection strategy, generational differences between artists had a strong-medium impact only for a few of them (13%), while in contrast artists' origin and nationality (82%), gender (79%), and color and skin tone (77%) were considered key factors for diversifying the lists.

In the combined task (*COMBO*), different factors seem to be interplaying. In the first pair *COMBO1* participants had a high agreement, and it is the only case where all the metrics and the aggregation mechanism agree on which list is more diverse. *COMBO2* represents an anomaly among the sets proposed. Indeed, both lists were formed by tracks similarly varied in terms of music features (no

---

[8]https://scikit-learn-extra.readthedocs.io/en/latest/modules/cluster.html#k-medoids

significant difference in terms of LL or HL features), but a list was formed by four tracks by white male EM artists versus a list formed by four tracks by white not-male artists. As result, in *COMBO2* lists were similarly different in terms of artists' attributes, and according to the Goodall metric their difference was not significant. However, for most of the participants (72%) the list with not-male artists was perceived as being more diverse than the one with only male artists. In the case of the *COMBO3* and *COMBO4* task, the low agreement corresponds to mixed results of the aggregation mechanisms, in line with the results found in the *TRACK* and *ARTIST* task.

When asked to indicate which among artist and track diversity they considered most in the last task, half of the participants indicated that they based their choices only or mostly on track diversity (50%), few participants only or mostly on artist diversity (13%), and the rest affirmed to consider both equally (37%).

A limitation of the results proposed until now derives from the use as a yardstick of the percentage agreement among participants. Indeed, such estimation does not take into account the agreement

Table 1. Percentage of users indicating which list was "most diverse" and pairwise distances d_i of each of the two lists. Two aggregation methods are shown: $\min(d_i)$ is the minimum distance within a list, and $\text{avg}(d_i)$ is the arithmetic average of the distances. For every list, we report the participants' agreement (%agree). For every pair, larger distance values are in **bold**, while the list with greater participant agreement is in *italic*. For a detailed description of the tasks see Section 3.

| Task | Chosen list (%agree) | $\min(d_i)$ | | | $\text{avg}(d_i)$ | | |
		$cosine_{HL}$	$cosine_{LL}$	Goodall	$cosine_{HL}$	$cosine_{LL}$	Goodall
TRACK1	A (7%)	.03	.03	-	.08	.05	-
	B (89%)	**.12**	**.10**	-	**.30**	**.19**	-
TRACK2	*A (85%)*	**.16**	**.07**	-	**.44**	**.27**	-
	B (10%)	.05	.07	-	.14	.16	-
TRACK3	*A (76%)*	.03	**.15**	-	**.26**	**.28**	-
	B (16%)	**.09**	.10	-	.17	.22	-
TRACK4	A (30%)	**.12**	.05	-	.20	.20	-
	B (58%)	.10	**.14**	-	**.29**	**.23**	-
ARTIST1	A (3%)	-	-	.41	-	-	.60
	B (94%)	-	-	**.79**	-	-	**.91**
ARTIST2	A (9%)	-	-	.41	-	-	.60
	B (86%)	-	-	**.63**	-	-	**.79**
ARTIST3	*A (70%)*	-	-	**.66**	-	-	**.76**
	B (20%)	-	-	.32	-	-	.72
ARTIST4	A (15%)	-	-	**.67**	-	-	.81
	B (78%)	-	-	.66	-	-	**.88**
COMBO1	A (19%)	.04	.08	.54	.13	.11	.66
	B (69%)	**.09**	**.19**	**.61**	**.28**	**.30**	**.86**
COMBO2	A (18%)	**.11**	.06	.40	**.39**	.18	.60
	B (72%)	.02	**.14**	**.42**	.25	**.24**	.60
COMBO3	A (43%)	**.09**	.08	**.58**	.16	.17	**.79**
	B (50%)	.01	**.17**	.26	**.27**	**.21**	.71
COMBO4	*A (55%)*	**.13**	**.14**	.35	**.42**	**.35**	.54
	B (33%)	.09	.04	**.65**	.19	.17	**.86**

that would be expected by chance, hence the final results may be overestimated [35]. To cope with this issue, in the following section we present the obtained results using reliability analysis.

5.3 Reliability Analysis

We use Cohen's kappa [16] to estimate the pairwise agreement between a participant and what the most diverse list should be according to the metrics (metric reliability – MR). We then report the average over each participant group, as described in Section 5.1.1. Similarly, we measure inter-rater reliability (IRR) by computing the pairwise agreement between participants using Cohen's kappa, then averaged to provide a single value. This value is equivalent to the Light's kappa [35], and we report in Table 2 and Table 3 such kappa for the MR and IRR respectively. In addition to kappa, we also report the 95% confidence interval obtained by bootstrapping [20]. The interpretation of kappa has been often debated, and there is no universally accepted scale [49]. For the sake of clarity, we refer to Cohen's suggested scale while interpreting the results obtained: values ≤ 0 indicate no agreement; .01−.20 none to slight; .21−.40 fair; .41−.60 moderate; .61−.80 substantial; .81−1.00 almost perfect agreement.

Results from the MR analysis of the *TRACK* task indicate how much the participants' perceptions agree with the diversity estimated with the cosine distance using the audio track features. We observe three major findings: 1) participants with low domain knowledge (G1) agree more in comparison to participants with high domain knowledge (G3); 2) participants in the WEIRD group (G4) agree more in comparison with one in the non-WEIRD group (G5); 3) participants with age between 18-35 (G6) agree more than participants with age over 35 (G7). Generally, the agreement in this task ranges from fair to substantial, almost perfect in very few cases.

Table 2. Metric reliability (MR) results. The Light's kappas and 95% confidence interval are reported separately for each group and each task. Values are in **bold**(*italic*) when they are 0.05 higher(lower) than the kappa measured across all participants. "T" refers to the *TRACK* task, "A" to the *ARTIST* task, "C" to the *COMBO* task. For details on the tasks see Section 3. G1, G2, G3 are groups of users respectively with low, medium, and high domain knowledge; G4 is WEIRD and G5 non-WEIRD; G6 is ≤ 35 years old, G7 is > 35 years old. For details on the groups see Section 5.1.1.

	T	T1-2	T3-4	A	A1-2	A3-4	C	C1-2	C3-4
G1	.63±.02	.81±.02	.46±.03	.64±.02	.87±.02	.57±.03	.25±.03	.56±.02	*.09± .04*
G2	.65±.02	*.72±.03*	**.60**±.04	.57±.04	.85±.02	.59±.04	.26±.03	.56±.04	**.20**±.05
G3	*.48±.03*	.73±.03	*.27±.05*	**.68**±.03	.83±.02	**.66**±.03	.20±.03	.50±.03	.17±.03
G4	.61±.01	.80±.01	.44±.02	.63±.01	.83±.01	.61±.01	.22±.01	.54±.01	*.10±.01*
G5	*.51±.03*	*.69±.04*	*.33±.05*	*.53±.04*	.83±.03	*.47±.05*	.28±.04	*.42±.03*	**.26**±.05
G6	.63±.01	.81±.01	.47±.02	.61±.01	.86±.01	.55±.02	.25±.01	.54±.02	.17±.02
G7	.55±.02	.74±.02	*.38±.02*	.59±.02	.81±.02	.60±.03	.22±.02	.50±.02	.10±.03
all	.59±.01	.77±.01	.43±.01	.60±.01	.83±.01	.56±.01	.24±.01	.52±.01	.15± .01

For the *ARTIST* task, we observe a different behaviour than the former task when comparing the perceived diversity with the diversity estimated by means of the Goodall metric. Indeed, in this case the domain knowledge is not so influential for determining the artist diversity, being not the task ratings based on any music-related aspect. Still, WEIRD and younger participants seem to agree more in comparison to their complementary groups. For this task, we observe moderate to almost perfect agreement for every analyzed group.

Table 3. Inter-rater reliability (IRR) results. The Light's kappas and 95% confidence interval are reported separately for each group and each task. Kappas are in **bold**(*italic*) when higher(lower) than 0.05 as to the kappa measured across all participants. "T" refers to the *TRACK* task, "A" to the *ARTIST* task, "C" to the *COMBO* task. For a detailed description of the groups see Section 5.1.1, for the tasks see Section 3.

	T	*T1-2*	*T3-4*	*A*	*A1-2*	*A3-4*	*C*	*C1-2*	*C3-4*
G1	**.45**±.02	**.66**±.03	.27±.03	.41±.03	**.76**±.03	.34±.03	.09±.01	.30±.03	*.03*±.01
G2	**.44**±.03	*.53*±.04	**.40**±.03	.37±.03	.72±.04	**.45**±.03	.14±.02	**.33**±.04	.09±.02
G3	*.30*±.02	.55±.03	*.10*±.03	**.46**±.03	.69±.04	**.43**±.04	.06±.02	.24±.03	**.17**±.02
G4	.41±.01	.64±.01	.23±.01	.41±.01	.70±.01	**.39**±.01	.10±.01	.28±.01	.09±.01
G5	*.32*±.03	*.50*±.04	.20±.04	*.30*±.04	.69±.05	.27±.04	.09±.02	*.17*±.03	**.15**±.03
G6	**.44**±.01	**.66**±.01	.26±.01	.38±.02	.73±.02	.33±.02	.11±.01	.30±.02	.12±.01
G7	*.33*±.02	*.56*±.03	.19±.02	.38±.02	.67±.03	**.39**±.02	.09±.01	.24±.02	.06±.01
all	.39±.01	.61±.01	.23±.01	.38±.01	.70±.01	.34±.01	.10±.00	.27±.01	.10±.01

Regarding the *COMBO* task, we observe an overall decrease of agreement level, expected as the complexity of this task increases in comparison to the former two tasks. Only in the case of a generational difference (G6 versus G7), we continue to see a greater agreement for the youngest participants. Looking at all the tasks, the kappa values for the pairs designed for having a more substantial difference (1-2) are greater than for the pairs where the difference between lists was slighter (3-4), confirming the task selection strategy presented in Section 3.4. Lastly, we notice that results from G4 (WEIRD) and G6 (≤ 35 years old) groups are almost in line with the results for all participants, being the majority of the subjects of our study belonging to such demographic groups.

The IRR analysis gives us a similar picture of the MR analysis, as shown in Table 3. Indeed, the agreement in the *TRACK* task is higher for participants with low domain knowledge, belonging to the WEIRD group, and aged between 18 and 35. Again, such difference is not present in the *ARTIST* task when comparing groups with different domain knowledge, while it is found comparing groups with different demographics. The inter-rater agreement in the *COMBO* task decreases, as observed in the MR analysis, and also the difference between pairs 1-2 and 3-4 is confirmed for the three tasks. In general, the inter-rater agreement can be interpreted as fair to moderate in the *TRACK* and *ARTIST* task, with few exceptions where it is substantial. For the *COMBO* task, in most of the cases there is none to slight inter-rater agreement.

6 DISCUSSION AND LIMITATIONS

In this section, we discuss our observations and try to answer the research questions about how different elements may influence perceived diversity. We also describe several limitations of this study.

With respect to RQ1 — *To what extent tracks' audio features and artists' attributes can be used to assess the perceived diversity?* — we have shown that audio-based features have some applicability for determining diversity, particularly when the case is clear enough that most people would agree that one list is more diverse than another. For this task, HL features may have an advantage over LL features in terms of interpretability and semantic meaning. However, LL features may be more discriminatory in assessing the diversity of a list, being perceptually less prone to subjective assessment in comparison to HL (e.g. differences in BPM are less subjective than differences in the happiness of a song), and being the HL constructs conceptually problematic by themselves (e.g. the

definition of what a "happy song" is) [44]. Artists' attributes can be used to model computationally a degree of diversity, which results from the reliability analysis confirmed to be shared among people from the same social group. When mixing different diversity dimensions (from tracks and artists), the increased complexity of the task results in a decreased agreement.

With respect to RQ2 — *To what extent domain knowledge and familiarity influence participants' perceptions of diversity?* — domain knowledge can be construed in multiple ways, including familiarity with artists or tracks. These are somewhat correlated, and all of them influence perceptions of diversity. For instance, self-declared music taste diversity within a music genre (electronic music) is correlated with familiarity and interacts with the perceptions of diversity. Moreover, track diversity perceptions strongly depend on domain knowledge and familiarity. Indeed, listeners with low domain knowledge are more easily satisfied with the diversity of a set of tracks and tend to agree more among them, while listeners with high domain knowledge are less easily satisfied in this regard and tend to agree less among them. Whilst these results might be counterintuitive at first, they agree with observations from previous work where it has been shown that experts tend to agree less on other types of music-related assessment, such as the emotion perceived while listening to music [64]. On the contrary, domain knowledge seems to have a limited effect on artist diversity compared to its effects over track diversity.

Limitations of the presented study are several and discussed hereafter. The generalization of the findings is limited by at least two factors. First, the participant population is strongly biased towards WEIRD societies, and consequently, results cannot provide a global validity. At the very least, a wider sample of participants coming from non-Western countries would be necessary to generalize our findings. Second, the choice to focus on electronic music has been motivated by the goal of assessing the relationship between domain knowledge and perceived diversity in a controlled environment. However, to validate our hypothesis, experimentation with other music genres should be foreseen.

A limitation in the design of the experiment is to explicitly ask participants to not consider popularity, likeness, or familiarity with the tracks and artists while assessing the diversity of the lists. Indeed, those aspects may interplay with the perception of diversity, and it may be hard to exclude them during the assessment. Further investigation is required to understand their role and relationship with diversity. In addition, listeners' characterization can be extended considering other music-related attributes, such as musical sophistication [53], but also considering other types of signals, such as personal listening histories to infer music tastes. Being the participation in our study voluntarily, it has been designed to limit the time needed to complete the survey to 30-35 minutes. This also determined the small number of sets for which participants have been asked to provide their opinion. Recruiting participants by providing an economic reward can be a method to overcome such issues. Lastly, the design of the study, the task and data selection, as well as computational aspects such as the features and metric selection cannot be immune to the authors' standpoint. We aim at improving these methods by confronting our approach with research communities from different fields.

7 CONCLUSION AND FUTURE WORK

Listeners, artists, and tracks, if imagined as *closed doors* may limit the understanding of cultural networks, in which a dual nature is embedded, where shared interests constitute social groups, but also shared publics form genres and styles [17]. Differences in these networks constitute what diversity is, and even if historically such differences have been often used for creating boundaries among cultures, for instance between Western music and others, the same differences can serve as a tool to promote and support multi- and intra-cultural environments [9, 15, 34, 66].

Through a user study, we analyzed how different diversity dimensions interact and how they are perceived, to prove that diversity cannot be understood if not considering its multifaceted nature. Even if features that are computable from the audio signal may help categorize music lists into coarse categories by diversity, complexities may arise while considering fine-grained ones. Consequently, such features may help generate diversity-aware recommendations to users with less domain knowledge but will be less useful for more specialized users. We showed that domain knowledge can be estimated by mixing self-declared answers and automatic scores, representing different aspects of familiarity and expertise within EM, and we provided a method for its assessment.

From another perspective, artists embed their diversity in their works and taking into account not only music features but also socially relevant characteristics can lead to new insights for designing diversity-aware music recommendations. The scenario described in this experiment represents a particular case of multistakeholder environment [1], where listeners represent the *consumers* of the content created by the artists, the *providers* in our settings. RS research in this area is recently emerging as a new sub-field wherein all the parts interacting are considered in the whole process of design, development, and evaluation of RS. Under this lens, our study of the characteristics of the several actors involved in Music RS may offer a new perspective for multistakeholder recommender systems diversity-related evaluation.

In order to cope with the limitations in terms of generalizability of the findings presented in this work, we plan to replicate the user study both considering a music genre different from electronic music and selecting a different participant population, not biased towards Western countries.

Reproducibility. To encourage reproducibility of the results, the code and data used in our study are publicly available.[9]

ACKNOWLEDGMENTS

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 127–158. https://doi.org/10.1007/s11257-019-09256-1

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/1498759.1498766

[3] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/3209978.3210024

[4] Richard Arneson. 2013. Egalitarianism. In *The Stanford Encyclopedia of Philosophy* (summer 2013 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[5] Emily M. Bender, Angelina McMillan-Major, Timnit Gebru, and Margaret Mitchell. 2021. On the dangers of stochastic parrots: can language models be too big?. In *FAccT 2021 - Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, Vol. 1. Association for Computing Machinery, 271–278. https://doi.org/10.1145/3442188.3445922 arXiv:2003.04628

---

[9]https://github.com/LPorcaro/music-diversity-analysis

[6] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. 2013. Essentia: An Audio Analysis Library for Music Information Retrieval. *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)* (2013), 493–498.

[7] Shyam Boriah, Varun Chandola, and Vipin Kumar. 2008. Similarity Measures for Categorical Data: A Comparative Evaluation. *Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics* 1 (2008), 243–254. https://doi.org/10.1137/1.9781611972788.22

[8] Georgina Born. 2020. Diversifying MIR: Knowledge and Real-World Challenges, and New Interdisciplinary Futures. *Transactions of the International Society for Music Information Retrieval (TISMIR)* 3 (2020), 193–204. https://doi.org/10.5334/tismir.58

[9] Georgina Born and David Hesmondhalgh. 2000. Introduction: On Difference, Representation and Appropriation in Music. In *Western Music and Its Others: Difference, Representation, and Appropriation in Music*. University of California Press, 1–58. http://www.ucpress.edu/book.php?isbn=9780520220843

[10] Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction* 12, 4 (2002), 331–370. https://doi.org/10.1023/A:1021240730564

[11] Sylvain Castagnos, Armelle Brun, and Anne Boyer. 2013. When Diversity Is Needed... But Not Expected!. In *International Conference on Advances in Information Mining and Management*. IARIA XPS Press, Lisbon, Portugal, 44–50. https://hal.inria.fr/hal-00931805

[12] Harr Chen and David R. Karger. 2006. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 429–436. https://doi.org/10.1145/1148170.1148245

[13] Li Chen, Wen Wu, and Liang He. 2013. How Personality Influences Users' Needs for Recommendation Diversity?. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 829–834. https://doi.org/10.1145/2468356.2468505

[14] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[15] Eric Clarke, Tia DeNora, and Jonna Vuoskoski. 2015. Music, Empathy and Cultural Understanding. *Physics of Life Reviews* 15 (2015), 61–88. https://doi.org/10.1016/j.plrev.2015.09.001

[16] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104

[17] Paul DiMaggio. 2011. Cultural Networks. *The Sage handbook of social network analysis* (2011), 286–310.

[18] Julia Driver. 2014. The History of Utilitarianism. In *The Stanford Encyclopedia of Philosophy* (winter 2014 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[19] Marina Drosou, H.V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in Big Data: A Review. *Big Data* 5, 2 (2017), 73–84. https://doi.org/10.1089/big.2016.0054

[20] Bradley Efron. 1992. *Bootstrap Methods: Another Look at the Jackknife.* Springer New York, New York, NY, 569–593. https://doi.org/10.1007/978-1-4612-4380-9_41

[21] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, California, USA) *(RecSys '14)*. Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/2645710.2645737

[22] Avriel Epps-darling, Romain Takeo Bouyer, and Henriette Cramer. 2020. Artist Gender Representation in Music Streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*.

[23] Katayoun Farrahi, Markus Schedl, Andreu Vall, David Hauger, and Marko Tkalčič. 2014. Impact of Listening Behaviour on Music Recommendation. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*. 1–6.

[24] Bruce Ferwerda, Mark P. Graus, Andreu Vall, Marko Tkalčič, and Markus Schedl. 2017. How Item Discovery Enabled by Diversity Leads to Increased Recommendation List Attractiveness. In *Proceedings of the Symposium on Applied Computing* (Marrakech, Morocco) *(SAC '17)*. Association for Computing Machinery, New York, NY, USA, 1693–1696. https://doi.org/10.1145/3019612.3019899

[25] Bruce Ferwerda and Marko Tkalčič. 2019. Exploring Online Music Listening Behaviors of Musically Sophisticated Users. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 33–37. https://doi.org/10.1145/3314183.3324974

[26] Bruce Ferwerda, Andreu Vall, Marko Tkalčič, and Markus Schedl. 2016. Exploring Music Diversity Needs Across Countries. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax, Nova Scotia, Canada) *(UMAP '16)*. Association for Computing Machinery, New York, NY, USA, 287–288. https://doi.org/10.1145/2930238.2930262

[27] Thomas B Fitzpatrick. 1988. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology* 124, 6 (1988), 869–871.

[28] Arthur Flexer. 2014. On Inter-rater Agreement in Audio Music Similarity. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*. 245–250.

[29] Arthur Flexer and Thomas Grill. 2016. The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. *Journal of New Music Research* 45, 3 (2016), 239–251. https://doi.org/10.1080/09298215.2016.1200631

[30] Arthur Flexer and Taric Lallai. 2019. Can We Increase Inter- and Intra-rater Agreement in Modeling General Music Similarity? *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)* (2019), 494–500.

[31] Ana Freire, Lorenzo Porcaro, and Emilia Gómez. 2021. Measuring Diversity of Artificial Intelligence Conferences. In *The AAAI 2021 Workshop on Diversity in Artificial Intelligence*. 1–8.

[32] Sreenivas Gollapudi and Aneesh Sharma. 2009. An Axiomatic Approach for Result Diversification. In *Proceedings of the 18th International Conference on World Wide Web* (Madrid, Spain) *(WWW '09)*. Association for Computing Machinery, New York, NY, USA, 381–390. https://doi.org/10.1145/1526709.1526761

[33] Juan Sebastián Gómez-Cañón, Estefanía Cano, Perfecto Herrera, and Emilia Gómez. 2020. Joyful for You and Tender for Us: The Influence of Individual Characteristics and Language on Emotion Labeling and Classification. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*.

[34] Line Grenier. 1989. From Diversity to Difference: the Case of Socio-Cultural Studies of Music. *New Formations* 9, Winter 1989 (1989).

[35] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology* 8, 1 (2012), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

[36] Natali Helberger. 2011. Diversity by design. *Journal of Information Policy* 1, 2011 (2011), 441–469. https://doi.org/10.5325/jinfopoli.1.2011.0441

[37] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The Weirdest People in the World? (2010). https://doi.org/10.2139/ssrn.1601785

[38] Geert H Hofstede, Gert Jan Hofstede, and Michael Minkov. 2005. *Cultures and organizations: Software of the mind.* Vol. 2. Mcgraw-hill New York.

[39] David Huron. 2004. Issues and Prospects in Studying Cognitive Cultural Diversity. In *Proceedings of the 8th International Conference on Music Perception & Cognition.*

[40] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. *AI Magazine* 41, 4 (2020), 79–95. https://doi.org/10.1609/aimag.v41i4.5312

[41] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[42] Christos Koutlis, Manos Schinas, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2020. GAP: Geometric Aggregation of Popularity Metrics. *Information* 11, 6 (2020). https://doi.org/10.3390/INFO11060323

[43] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in Recommender Systems – A Survey. *Knowledge-Based Systems* 123 (2017), 154–162. https://doi.org/10.1016/j.knosys.2017.02.009

[44] Cynthia C.S. Liem and Chris Mostert. 2020. Can't Trust the Feeling? How Open Data Reveals Unexpected Behavior of High-Level Music Descriptors. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*.

[45] Qika Lin, Yaoqiang Niu, Yifan Zhu, Hao Lu, Keith Zvikomborero Mushonga, and Zhendong Niu. 2018. Heterogeneous knowledge-based attentive neural networks for short-term music recommendations. *IEEE Access* 6 (2018), 58990–59000. https://doi.org/10.1109/ACCESS.2018.2874959

[46] Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50–60. https://doi.org/10.1214/aoms/1177730491

[47] Marti A. Hearst. 1999. User Interfaces and Visualization. In *Modern Information Retrieval*, Ricardo Baeza-Yates (Ed.). ACM Press, New York, NY, USA, Chapter 10, 257–340.

[48] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359174

[49] Mary L Mchugh. 2012. Interrater Reliability: the Kappa Statistic. *Biochemia Medica (Zagreb)* 22, 3 (2012), 276–282.

[50] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI EA '06)*. Association for Computing Machinery, New York, NY, USA, 1097–1101. https://doi.org/10.1145/1125451.1125659

[51] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and Inclusion Metrics in Subset Selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 117–123. https://doi.org/10.1145/3375627.3375832

[52] Jean Molino, J A Underwood, and Craig Ayrey. 1990. Musical Fact and the Semiology of Music. *Music Analysis* 9, 2 (jan 1990), 105–156. https://doi.org/10.2307/854225

[53] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. 2014. The Musicality of Non-musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE* 9, 6 (2014). https://doi.org/10.1371/journal.pone.0089642

[54] High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Online; accessed 7 July 2021].

[55] Sergio Oramas, Luis Espinosa-Anke, Mohamed Sordo, Horacio Saggion, and Xavier Serra. 2016. Information Extraction for Knowledge Base Construction in the Music Domain. *Data & Knowledge Engineering* 106 (2016), 70–83. https://doi.org/10.1016/j.datak.2016.06.001

[56] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. 2017. A Deep Multimodal Approach for Cold-start Music Recommendation. *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems - DLRS 2017* (2017), 32–37. https://doi.org/10.1145/3125486.3125492 arXiv:1706.09739

[57] Hae-Sang Park and Chi-Hyuck Jun. 2009. A Simple and Fast Algorithm for K-medoids Clustering. *Expert Systems with Applications* 36, 2, Part 2 (2009), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

[58] Minsu Park, Ingmar Weber, Mor Naaman, and Sarah Vieweg. 2015. Understanding Musical Diversity via Online Social Media. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM'15)*.

[59] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355. https://doi.org/10.1007/s11257-011-9115-7

[60] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

[61] Kyle Robinson, Dan Brown, and Markus Schedl. 2020. User Insights on Diversity in Music Recommendation Lists. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*.

[62] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are "Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 595–604. https://doi.org/10.1145/3331184.3331215

[63] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do User Preferences and Evaluation Measures Line Up?. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 555–562. https://doi.org/10.1145/1835449.1835542

[64] Markus Schedl, Emilia Gomez, Erika S. Trent, Marko Tkalcic, Hamid Eghbal-Zadeh, and Agustin Martorell. 2018. On the interrelation between listener characteristics and the perception of emotions in classical orchestra music. *IEEE Transactions on Affective Computing* 9, 4 (2018), 507–525. https://doi.org/10.1109/TAFFC.2017.2663421

[65] Markus Schedl, Emilia Gómez, and Julián Urbano. 2014. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends® in Information Retrieval* 8, 2-3 (2014), 127–261. https://doi.org/10.1561/1500000042

[66] Xavier Serra. 2011. A Multicultural Approach in Music Information Research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. 151–156.

[67] Amit Singhal. 2001. Modern information retrieval: a brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (2001), 2001.

[68] Daniel Steel, Sina Fazelpour, Kinley Gillette, Bianca Crewe, and Michael Burgess. 2018. Multiple diversity concepts and their ethical-epistemic implications. *European Journal for Philosophy of Science* 8, 3 (2018), 761–780. https://doi.org/10.1007/s13194-018-0209-5

[69] Andy Stirling. 2007. A General Framework for Analysing Diversity in Science, Technology and Society. *Journal of The Royal Society Interface* 4, 15 (2007), 707–719. https://doi.org/10.1098/rsif.2007.0213

[70] Scientific UN Educational and Cultural Organisation (UNESCO). 2001. UNESCO Universal Declaration on Cultural Diversity. http://portal.unesco.org/en/ev.php-URL_ID=13179&URL_DO=DO_TOPIC&URL_SECTION=201.html [Online; accessed 7 July 2021].

[71] Julián Urbano, Brian McFee, J. Stephen Downie, and Markus Schedl. 2012. How Significant is Statistically Significant? The Case of Audio Music Similarity and Retrieval. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*. 181–186.

[72] Julián Urbano, Markus Schedl, and Xavier Serra. 2013. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems* 41, 3 (2013), 345–369. https://doi.org/10.1007/s10844-013-0249-4

[73] Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-Oriented Diversity in Recommender Systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 1211–1212. https://doi.org/10.1145/2009916.2010124

[74] Olivia R. Ware, Jessica E. Dawson, Michi M. Shinohara, and Susan C. Taylor. 2020. Racial Limitations of Fitzpatrick Skin Type. *Cutis* 105, 2 (2020), 77–80.

[75] Hans W. Wendt. 1972. Dealing with a Common Problem in Social Science: A Simplified Rank-biserial Coefficient of Correlation Based on the U Statistic. *European Journal of Social Psychology,* 2, 4 (1972), 463–465. https://doi.org/10.1002/ejsp.2420020412

[76] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race and Power in AI. *AI Now Institute* April (2019).

[77] Martijn C Willemsen, Mark P Graus, and Bart P Knijnenburg. 2016. Understanding the Role of Latent Feature Diversification on Choice Difficulty and Satisfaction. *User Modeling and User-Adapted Interaction* 26, 4 (2016), 347–389. https://doi.org/10.1007/s11257-016-9178-6

[78] Qian Zhao, Gediminas Adomavicius, F. Maxwell Harper, Martijn Willemsen, and Joseph A. Konstan. 2017. Toward Better Interactions in Recommender Systems: Cycling and Serpentining Approaches for Top-N Item Lists. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1444–1453. https://doi.org/10.1145/2998181.2998211

[79] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05)*. Association for Computing Machinery, New York, NY, USA, 22–32. https://doi.org/10.1145/1060745.1060754

A   SUPPLEMENTARY MATERIAL
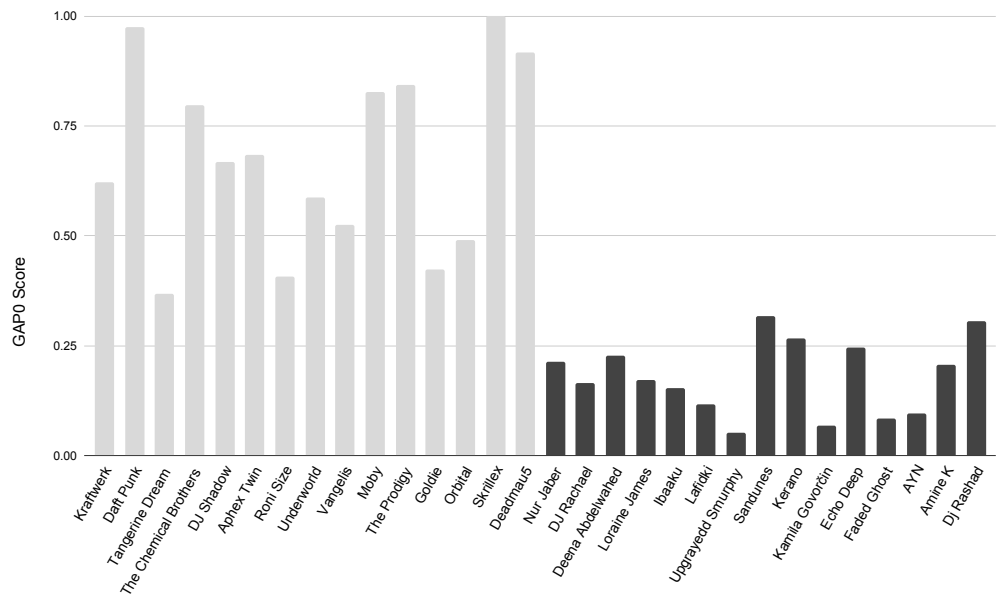

Fig. 3. List of artists and corresponding GAP0 score for assessing the participants' familiarity with EM artists. Artists selected within Allmusic most representative are in light grey (*left*), while artists selected within the ones shown in the survey are in dark grey (*right*). "Skrillex" results to be the most popular, whilst "Upgrayedd Smurphy" the one with less social presence.

Table 4. List of tracks included in the *TRACK* task. (* Following the controversy about the previous alias used by the artist ("The Black Madonna") we report here the new alias. Such event occurs during the time when authors collected participants answers. For more details: https://pitchfork.com/news/the-black-madonna-changes-name-to-the-blessed-madonna/).

Task	List	Artist Name	Track Name
TRACK1	List A	Resensed & Asterz	Last Time
		KELTEK & B-Front	Step Into The Game
		Hard Driver x Max Enforcer	Before I Wake
		Exilium	I Got All The Time
	List B	999999999	X0004000X
		HOSH & 1979 Feat. Jalja	Midnight (The Hanging Tree)
		Tibi Dabo	Komorebi
		Kid Francescoli	Moon
TRACK2	List A	Pretty Light	Yellow Bird
		Bonobo	Kong
		Technoboy	Ti sento
		Laurent Garnier	The man with the red face
	List B	Amelie Lens	The Future
		Nicole Moudaber	Own
		Russell E.L. Butler	Gulf Stream
		Satoshi Tomiie	Resonant
TRACK3	List A	The Blessed Madonna*	Stay
		Beatrice Dillon	Workaround One
		Miss DJAX	Headbangin'
		Helena Hauff	Spur
	List B	Nur Jaber	Until I Collapse
		DJ Rachael	Silaye
		Deena Abdelwahed	Ababab
		Nina Kraviz	Desire
TRACK4	List A	Echo Deep	Maasai Groove
		Amine K	Dar Gnawa
		Ricardo Villalobos	Lazer@Present
		Solomun	The way back
	List B	Aphex Twin	Windowlicker
		Dubfire	Roadkill
		Orieta Chrem	Delfines
		Dj Rashad	Feelin

Table 5. List of artists included in the *ARTIST* task. Artists' gender is reported as male (M), female (F), or non-binary (X). Skin tone is categorized following the Fitzpatrick scale [27]. Debut refers to the decade in which the artists started playing.

Task / List	Artist Name	Birthplace	Gender	Skin Tone	Debut
ARTIST1 / List A	Pretty Lights	United States	M	I	2000'
	Bonobo	United Kingdom	M	II	1990'
	Technoboy	Italy	M	II	1990'
	Laurent Garnier	France	M	II	1980'
List B	Amelie Lens	Belgium	F	II	2010'
	Nicole Moudaber	Nigeria	F	III	2000'
	Russell E.L. Butler	Bermuda	X	IV	2010'
	Satoshi Tomiie	Japan	M	II	1980'
ARTIST2 / List A	The Blessed Madonna*	United States	X	I	2010'
	Beatrice Dillon	United Kingdom	F	I	2010'
	Miss DJAX	The Netherlands	F	II	1980'
	Helena Hauff	Germany	F	II	2010'
List B	Nur Jaber	Lebanon	F	II	2010'
	DJ Rachael	Uganda	F	VI	1990'
	Deena Abdelwahed	Tunisia	X	III	2010'
	Nina Kraviz	Russia	F	I	2000'
ARTIST3 / List A	Loraine James	United Kingdom	X	V	2010'
	Ibaaku	Senegal	M	VI	2010'
	Lafidki	Cambodia	M	III	2010'
	Upgrayedd Smurphy	Mexico	F	I	2010'
List B	Sandunes	India	F	III	2010'
	Kerano	India	M	III	2010'
	Kamila Govorčin	Chile	F	III	2010'
	Nicolas Jaar	United States	M	II	2010'
ARTIST4 / List A	Echo Deep	South Africa	M	V	2010'
	Amine K	Morocco	M	IV	2000'
	Ricardo Villalobos	Chile	M	III	1990'
	Solomun	Bosnia-Herzegovina	M	II	2000'
List B	Aphex Twin	Ireland	M	I	1980'
	Dubfire	Iran	M	III	1990'
	Orieta Chrem	Peru	F	II	2000'
	Dj Rashad	United States	M	VI	1990'

Table 6. List of tracks and artists included in the *COMBO* task.

Task	List	Artist Name	Track Name
COMBO1	List A	Skrillex	Bangarang
		Haezer	Bass Addict
		Clean Tears	Bouquet
		Carl Nunes	Limitless
	List B	Le Bask	Hardchoriste
		Faded Ghost	Moon Rain
		Andres Gil	Scene
		AYN	Where It Went
COMBO2	List A	Pretty Light	Yellow Bird
		Bonobo	Kong
		Technoboy	Ti sento
		Laurent Garnier	The man with the red face
	List B	The Blessed Madonna*	Stay
		Beatrice Dillon	Workaround One
		Miss DJAX	Headbangin'
		Helena Hauff	Spur
COMBO3	List A	Nur Jaber	Until I Collapse
		DJ Rachael	Silaye
		Deena Abdelwahed	Ababab
		Nina Kraviz	Desire
	List B	Sandunes	Exit Strategy
		Kamila Govorčin	Daydream
		Nicolas Jaar	A time for us
		Kerano	Thirsty
COMBO4	List A	Ibaaku	Discuting Food
		DJ Rachael	Silaye
		Echo Deep	Maasai Groove
		Bosaina	Abalone on the Grass
	List B	Amelie Lens	The Future
		Aphex Twin	Windowlicker
		Ricardo Villalobos	Lazer@Present
		Peggy Gou	Starry Night

Table 7. List of features used for computing the distance using the cosine metric.

High-Level	danceability	mood sad
	mood aggressive	timbre
	mood happy	tonal / atonal
	mood party	voice / instrumental
	mood relaxed	
Low-Level	average loudness	spectral energy
	barkbands crest	spectral energyband high
	barkbands flatness db	spectral energyband low
	barkbands kurtosis	spectral energyband middle high
	barkbands skewness	spectral energyband middle low
	barkbands spread	spectral entropy
	dissonance	spectral flux
	dynamic complexity	spectral kurtosis
	erbbands crest	spectral rms
	erbbands flatness db	spectral rolloff
	erbbands kurtosis	spectral skewness
	erbbands skewness	spectral strongpeak
	erbbands spread	zero crossing rate
	hfc	chords changes rate
	melbands crest	chords number rate
	melbands flatness db	chords strength
	melbands kurtosis	hpcp crest
	melbands skewness	hpcp entropy
	melbands spread	tuning diatonic strength
	pitch salience	tuning equal tempered deviation
	silence rate 30dB	tuning frequency
	silence rate 60dB	tuning nontempered energy ratio
	spectral centroid	onset rate
	spectral decrease	BPM

Table 8. Report of Mann–Whitney U test. The results for every comparison presented to participants are shown, separately for the cosine metric computed using HL ($cosine_{HL}$) and LL ($cosine_{LL}$) features, and the Goodall metric computed with the artists' attributes. Together with the median distance for every list, the U-statistics, the significance at $p = 0.05$ and $p = 0.1$, and the effect size (computed using the Wendth formula [75]) are reported. For every pair, the list with the highest median is in **bold**.

	Task	Median(d_i) List A	Median(d_i) List B	U	Significance at p=0.05	Significance at p=0.1	Effect Size
$cosine_{HL}$	TRACK1	.08	**.30**	1	True	True	.94
	TRACK2	**.50**	.14	3	True	True	.83
	TRACK3	**.28**	.13	10	False	False	.44
	TRACK4	.17	**.33**	9	False	False	.5
	COMBO1	.13	**.28**	8	False	False	.56
	COMBO2	**.44**	.28	9	False	False	.5
	COMBO3	.13	**.27**	12	False	False	.33
	COMBO4	**.43**	.19	6	False	True	.67
$cosine_{LL}$	TRACK1	.05	**.18**	0	True	True	1
	TRACK2	**.31**	.16	5	True	True	.72
	TRACK3	**.27**	.20	12	False	False	.33
	TRACK4	**.23**	.22	15	False	False	.17
	COMBO1	.10	**.32**	0	True	True	1
	COMBO2	.20	**.21**	13	False	False	.27
	COMBO3	.16	**.20**	10	False	False	.44
	COMBO4	**.38**	.17	4	True	True	.78
Goodall	ARTIST1	.68	**.90**	0	True	True	1
	ARTIST2	.60	**.82**	5	False	True	.72
	ARTIST3	**.80**	.79	17	False	False	.03
	ARTIST4	.82	**.91**	11	False	False	.39
	COMBO1	.83	**.89**	9	False	False	.5
	COMBO2	**.63**	.60	15	False	False	.17
	COMBO3	**.83**	.78	17	False	False	.05
	COMBO4	.57	**.88**	0	True	True	1