Assessing the Impact of Music Recommendation Diversity on Listeners: A Longitudinal Study

LORENZO PORCARO and EMILIA GÓMEZ, Music Technology Group, UPF, Spain and Joint Research Centre, European Commission
CARLOS CASTILLO, Web Science and Social Computing Group, UPF, Spain and ICREA, Spain

We present the results of a 12-week longitudinal user study wherein the participants, 110 subjects from Southern Europe, received on a daily basis Electronic Music (EM) diversified recommendations. By analyzing their explicit and implicit feedback, we show that exposure to specific levels of music recommendation diversity may be responsible for long-term impacts on listeners' attitudes. In particular, we highlight the function of diversity in increasing the openness in listening to EM, a music genre not particularly known or liked by the participants previous to their participation in the study. Moreover, we demonstrate that recommendations may help listeners in removing positive and negative attachments towards EM, deconstructing pre-existing implicit associations but also stereotypes associated with this music. In addition, our results show the significant influence that recommendation diversity has in generating curiosity in listeners.

CCS Concepts: • **Information systems** → **Information retrieval**; • **Human-centered computing** → **User studies**; • **Applied computing** → **Sound and music computing;**

Additional Key Words and Phrases: Music information retrieval, longitudinal analysis, long-term impact

ACM Reference format:
Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2024. Assessing the Impact of Music Recommendation Diversity on Listeners: A Longitudinal Study. *ACM Trans. Recomm. Syst.* 2, 1, Article 3 (March 2024), 47 pages.
https://doi.org/10.1145/3608487

1 INTRODUCTION

Recommender Systems (RS) affect several choices in our daily life, helping us choose, for instance, the news we read, the movies we watch, the job positions we apply for, or the music we listen to. Besides the end-users consuming the recommendations, RS also affect those producing the items being recommended: Artists' royalty revenues may depend on whether they are recommended or not within a streaming platform; the fame of a brand in e-commerce may rely upon

This work is part of the project Musical AI - PID2019-111403GB-I00/AEI/ 10.13039/501100011033 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI). This work is also partially supported by the HUMAINT programme (Human Behaviour and Machine Intelligence), Joint Research Centre, European Commission. The project leading to these results received funding "la Caixa" Foundation (ID 100010434), under agreement LCF/PR/PR16/51110009, an from EU-funded projects "SoBigData++" (grant agreement 871042) and "FINDHR" (grant agreement 101070212).
Authors' addresses: L. Porcaro, Via Enrico Fermi, 2749, 21027 Ispra, Italy; email: lorenzo.porcaro@ec.europa.eu; E. Gómez, Calle Inca Garcilaso, 3, 41092 Seville, Spain; email: emilia.gomez-gutierrez@ec.europa.eu; C. Castillo, Carrer de Tànger, 122, 08018 Barcelona, Spain; email: carlos.castillo@upf.edu.

2770-6699/2024/03-ART3 $15.00
https://doi.org/10.1145/3608487

if its products are displayed or not as similar to the ones previously purchased by users; the time spent unemployed may depend on whether a profile is shown or not to potential employers. These are just examples of the several RS stakeholders subject to disparate and sometimes unintended impacts [42].

The awareness of these impacts is at the basis of the flourishing fair ranking and recommendations literature [71]. For instance, Hasan et al. [35] show that RS potentially increases excessive video usage on online platforms. Adomavicius et al. [1] highlight the role of recommendations in manipulating consumers' preference ratings. Fabbri et al. [20] investigate the role of RS in promoting users' radicalization. Notwithstanding, RS may certainly be responsible also for positive changes. Hauptmann et al. [36] propose an app for healthy food recommendations that positively affect nutritional behavior. Music recommendations have been used for helping recover the musical memory of people with Alzheimer's disease [66]. In the work by Starke et al. [85] users' adoption of energy-saving measures is boosted, thanks to the recommendation interface design. Other examples are provided in the overview of RS stakeholders, values, and risks by Jannach and Bauer [40].

Among the recommendation characteristics under the spotlight in RS impact-oriented research, diversity has drawn the interest of researchers, practitioners, policy-makers, and also affected communities because of its latent influence on individuals' choices. In particular, *exposure diversity* [38] as mediated by recommender systems is a research topic attracting scholars from different disciplines, especially in relation to its impact on human rights such as inclusion, non-discrimination, and fairness. Previous works in the RS literature have investigated how recommendations may influence aspects such as consumption or sales diversity, sometimes focusing on the impact of diversity on other recommendation characteristics, such as their usefulness or attractiveness, e.g., Reference [98]. Nevertheless, in this strain of RS research, longitudinal user studies are still quite rare.

We contribute to this latter corpus of research with this work, the first longitudinal user study in the Music RS literature presenting an analysis of the impact of music recommendation diversity on listeners' attitudes, specifically on: (1) openness in listening, (2) willingness to discovery, (3) implicit association, and (4) stereotyping of an unfamiliar music genre. It consists of a 12-week study wherein the participants, 110 subjects from Southern Europe, received daily **Electronic Music (EM)** track recommendations with different levels of diversity. We propose a critical evaluation that takes into account insights from the music psychology and education fields. Indeed, music exposure is proven to have the power of reducing stereotypes and prejudice against unknown or unfamiliar cultures. In addition, in the music domain, repeated exposure and familiarity have been linked to the construction of aesthetic preferences. Under this lens, we evaluate the impact of recommendation diversity not from a behavioral perspective but instead assessing how music recommendations may be a vehicle for attitudinal change. Specifically, we aim at answering the following questions:

[RQ1] To what extent can listeners' implicit and explicit attitudes towards an unfamiliar music genre be affected by exposure to music recommendations?

[RQ2] What is the relationship between music recommendation diversity and the impact on listeners' attitudes?

By analyzing participants' explicit and implicit feedback, we show that exposure to specific levels of music recommendation diversity may be responsible for long-term impacts on listeners' attitudes. In particular, we highlight the function of diversity in increasing the openness in listening to EM, a music genre not particularly known or liked by the participants previous to their participation in the study. Moreover, we demonstrate that recommendations may help listeners

in removing positive and negative attachments towards EM, deconstructing pre-existing implicit associations but also stereotypes associated with this music. In addition, our results show the significant influence that recommendation diversity has in generating curiosity in listeners.

The rest of the article is organized as follows: Section 2 starts with an overview of impact assessment practices, followed by a brief survey of recent developments of impact-oriented RS diversity research, and, last, presents several works on the impact of music exposure on listeners. Afterwards, Section 3 describes the user study design, while Section 4 reports the process to create the music recommendation to which study participants have been exposed. Then, Section 5 presents the results of our analysis, discussed together with their limitations in Section 6. Based on our findings, we examine future design implications for Music RS in Section 7 to finally draw conclusions in Section 8.

2 BACKGROUND AND RELATED WORK

Algorithmic Impact Assessment (AIA) is a complex process that goes beyond the development of practices to measure quantitatively some kind of change. Instead, it includes the involvement of several actors, starting from the system designers arriving at the community affected by algorithmic systems. In Section 2.1, we introduce the concept of *impact assessment*, in particular focusing on recent proposals of AIA. Afterwards, in Section 2.2, we center our attention on RS simulation-based frameworks, which have attracted the attention of the practitioners interested in assessing the impact of these systems. Finally, in Section 2.3, we present several insights from the music psychology field on the impact of music exposure on listeners.

2.1 Algorithmic Impact Assessment

In its simplest form, **Impact Assessment (IA)** can be defined as "[...] the process of identifying the future consequences of a current or proposed action. The *impact* is the difference between what would happen with the action and what would happen without it" [39]. In this area, AIA is relatively a new field in comparison to other frameworks such as ***Environmental Impact Assessment* (EIA)** [63], ***Social* (and *Societal*) *Impact Assessment* (SIA)** [51, 92, 93], ***Human Rights Impact Assessment* (HRIA)** [47], or ***Cultural Impact Assessment* (CIA)** [70]. Nevertheless, its importance is quickly growing.

AIA can be defined as a set of "emerging governance practices for delineating accountability, rendering visible the harms caused by algorithmic systems, and ensuring practical steps are taken to ameliorate those harms" [62]. A few aspects of AIA are addressed in detail below, but we point the reader interested in a comprehensive overview and discussion on current AIA practices towards the reports published by the non-profit organisations *AI Now Institute* [76] and *Data & Society* [65].

As discussed by Vecchione et al. [94] in the context of algorithmic auditing, most of the impact that may result from the interaction with an algorithmic system appears beyond discrete moments of decision-making. This is particularly true for RS, wherein the impact may not be evident after a single interaction, but instead be the fruit of multiple exposures through time. Second, as Metcalf et al. [62] argue, the impact is co-constructed by all the actors linked to an algorithmic system: developers, designers, decision-makers, public and private organizations, and, most importantly, the affected communities. Therefore, it is fundamental to involve each of these actors while defining the AIA practice, a vision shared also in Reference [94].

While the exploration of the long-term impact is already on the agenda of RS practitioners, the involvement of the wider community of people affected by RS is still rare to find. In the music field, a notable exception is the work by Ferraro [21] considering the artists' perspective on the impact of music recommendation. Another example that is worth mentioning is the work done in the Algorithmic Responsibility research area at Spotify [84].

2.2 Diversity in Impact-oriented Recommender Systems Research

As recently observed by Liang and Willemsen [56], longitudinal studies are not common in RS literature and their recent work is a notable exception, as also the study presented by Hauptmann et al. [36] presenting a user study on a nutrition assistance recommender system. On the contrary, the interest in impact-oriented RS research using synthetic data and simulation environments is rapidly growing within and outside the RS community [18]. However, this growing interest is accompanied by an equally growing concern about the high heterogeneity and low transparency of methods, evaluation practices, and more general assumptions on which such simulation studies are built [99]. Next, we review the simulation-based recommender system literature wherein the impact of recommendations on different kinds of diversity has been considered.

Agent-Based Modelling (ABM) has been applied in several works to understand the long-term impact of RS [2]. For instance, Zhang et al. [103] center their attention on simulating users' consumption strategies, showing how, relying on recommendations, users may end up in the long-term contributing to the decrease of aggregate diversity. Zhou et al. [104] use ABM to study how *preference bias*—the distortion in users' self-reported ratings caused by recommendations—influences the performance of recommender systems. In particular, they show how the bias introduced into the system through the users' ratings may negatively influence the overall diversity of the recommended items.

Further examples of simulated environments can be found in the RS literature analyzing the impact of *feedback loops*. Mansoury et al. [59] design an iterative model to analyze the feedback loop, showing how it may cause a decline in aggregate diversity. Moreover, Jiang et al. [43] provide a theoretical analysis of the relationship between feedback loops, echo chambers, and filter bubbles. Instead, Chaney et al. [12] by simulating different models of users' engagement with RS prove the impact of feedback loops on the homogenization of users' behaviors. Similarly, Aridor et al. [6], by using numerical simulations to model user decision-making processes, provide an explanation of the findings of a previous study by Nguyen et al. [67] on the influence of recommender systems on content diversity. In the latter, the authors found that users' interacting with the provided recommendations ended up consuming more diverse content in comparison to the users who did not. Aridor and colleagues confirm such results but also observe an increase in the homogeneity across users, i.e., decrease in aggregate diversity. Content diversity is also considered by Anderson et al. [5], observing a connection between recommendations and long-term reduction of diversity.

Under a different lens, *sales diversity*, defined as the concentration of consumer purchases, is at the center of attention in a series of studies by Fleder and Hosangar [24, 25] and Lee and Hosanagar [53, 54]. Fleder and Hosangar prove that collaborative filtering recommender systems exert a *concentration bias*, early delineating the idea that it may be possible that, at the individual level, diversity may increase, while at the aggregate level diversity may decrease. Lee and Hosanger, using field experiment data, confirm such results, highlighting how users do effectively explore novel items, thanks to the recommendations, but such explorations are highly correlated among users.

Another strain of simulation-based analysis focuses on comparing the performance of different typologies of recommender systems. Hazrati et al. [37] create a simulated environment where users are exposed to different kinds of recommender systems, proving that non-personalized methods produce the lowest diversity in terms of users' choices. Jannach et al. [41] focus on the analysis of recommendation techniques using an iterative approach, wherein users are assumed to interact with a specific fraction of the recommended items. The authors show that, in terms of the spread and coverage of recommendations, several systems analyzed led to an increased concentration over time. Using the same methodology but specifically on session-based recommender systems, Ferraro et al. [23] find similar results in terms of spread and coverage.

A limitation of current impact-oriented RS research is the lack of an assemblage of a wide range of expertise. Indeed, being RS research mostly driven by computer science-inspired approaches, most of what until now has been measured as impact is the result of technical and engineering knowledge. Despite the several efforts to properly develop robust metrics and evaluation procedures, the narrowness of the considered approaches has limited the concept of impact itself to what is measurable according to such procedures, a limit that should be overcome to make algorithmic impact assessment practices shared and effective.

2.3 Impact of Music Exposure

Several scholars in the music psychology and education field investigated the role of music exposure in influencing stereotypes and attitudes, starting from the idea that repeated exposure to a stimulus object may affect the preference towards such object, the so-called *mere exposure effect* theorized by Zajonc [102]. We highlight some examples hereafter.

Greitemeyer and colleagues [32] prove that exposure to music with pro-integration lyrics may reduce prejudice and discrimination towards immigrant groups, and later in Reference [31] they show that exposing listeners to music with pro-equality lyrics may enhance positive attitudes and behaviors toward women. In both cases, the authors found that the musical characteristics of the music exposed and the preference for it do not influence such impacts. Clarke et al. [14] investigate the relationship between music, empathy, and cultural understanding, showing that music exposure may indeed generate a sense of affiliation with unknown cultures. Their results are expanded in Reference [96], confirming that, especially in listeners with high trait empathy, music may increase positive implicit attitudes towards images representing members of foreign cultures. Tu [91] provides empirical evidence that exposing young students to 10 minutes of a Chinese music curriculum, when prolonged for 10 weeks, may impact the attitudes towards Chinese people. In a similar study, Sousa et al. [82] show that exposure to Cape Verdean songs together with Portuguese songs may reduce anti-dark-skinned stereotyping among light-skinned Portuguese children. Even if these examples consider different stimuli and different subjects, they all provide evidence that music listening may influence the idea that we have about other social groups, and more broadly about other cultures.

Another strain of research in the field of Music Psychology has been interested in understanding the impact of repeated exposure on music preference, based on Berlyne's psychobiological theory (see Reference [13] for an overview). Among his several contributions, he theorized the existence of a relationship between aesthetic preferences and familiarity. Even if some studies support his claims, e.g., Reference [89], while others confute it, e.g., Reference [58], nowadays it is commonly accepted that familiarity with music has a prominent role in the development of preferences [44], a topic in which also neuroscience practitioners extensively debate [26]. In the Music RS literature, Sguerra et al. [80], performing large-scale analysis of users' music consumption on Deezer, prove the validity of Berlyne's theory by focusing on music discovery, showing how users' interests may evolve in relation to the repeated exposure. These studies motivate our interest in exploring the impact that music recommendation diversity may have, connecting exposure, familiarity, and preferences. Indeed, while music recommender systems are known to be influential on exposure diversity [38, 73], little is known about how such exposure diversity may affect the users' opinions, beliefs, and attitudes.

3 STUDY DESIGN

The study is divided into three main stages, namely, *PRE*, *COND*, and *POST*, preceded by a prescreening stage, detailed hereafter. First, we designed an online survey to select subjects matching a set of criteria, presented in Section 3.2. Afterwards, we started collecting participants' data using

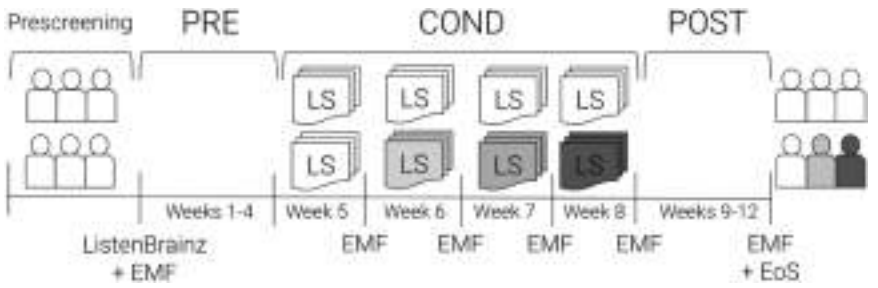

Fig. 1. High-level view of the longitudinal study. Participants are divided into two groups: low-diversity recommendations (*white*) and high-diversity recommendations (*shaded grey*). LS stands for Listening Sessions. EMF stands for Electronic Music Feedback questionnaire. EoS stands for End-of-Study survey.

two methods. First, we asked them to create a ListenBrainz[1] account to gather information about their listening habits (Section 3.3). Additionally, participants completed the **Electronic Music Feedback (EMF)** questionnaire, where they provided their opinion about several aspects of EM (Section 3.4). For the following four weeks, no further actions were required to the participants. This stage of the experiment is referred to as *PRE*.

In the 5th week, participants started to be exposed to music recommendations in what we call the *COND* (conditioning) stage. At that point, participants were already randomly divided into two groups, one receiving **high diversity (HD)** and the other **low diversity (LD)** recommendations, created following the procedure described in Section 4. For four weeks, from Monday to Friday, participants received on a daily basis an audio mix to be listened to, for a total of 20 listening sessions (Section 3.5). After each listening session, additional feedback was collected by asking participants their impressions about the music listened to. During this stage, they were also asked to complete the EMF questionnaire on a weekly basis, on Saturday. At the start of the 9th week, the *COND* stage ended and the *POST* stage started. Again for four weeks, no further actions were required to the participants. Finally, at the end of the 12th week, we asked participants first to fill for the last time the EMF questionnaire and then to fill out the **End-of-Study (EoS)** survey described in Section 3.6. Figure 1 depicts the study's high-level structure.

The study temporal design has been motivated by the following reasons: First, the four initial weeks in the *PRE* stage have been necessary to collect enough data to understand the participants' listening habits before starting the conditioning stage. Second, the choice of short measurement intervals weekly based in the *COND* stage has made it possible to analyze participants' responses with regard to the number of recommendations to which they have been exposed. Last, the last four weeks waited before to perform again the measurements in the *POST* stage have made it possible to separate the direct impact of the conditioning stage from the effect in the long-term. While different temporal designs could lead to possibly different results, the absence of longitudinal studies in the field of Music RS makes it hard to compare with prior literature.

3.1 Recruitment and Informed Consent

The **Institutional Committee for Ethical Review of Projects (CIREP)** at Universitat Pompeu Fabra approved the study design and confirmed the compliance of the research project with the data protection legal framework, namely, with the **European General Data Protection Regulation (EU) 2016/679 (GDPR)** and Spanish Organic Law 3/2018, of December 5th, on **Protection of Personal Data and Guarantee of Digital Rights (LOPDGDD)**. A digital copy of the

[1] https://listenbrainz.org

submitted documentation, the ethics certificate, and the data protection certificate are available upon request.

Before the main study, a smaller-scale pilot study was conducted to test the data collection process. The pilot study took place in February and April 2022, and the main study took place from May to July 2022. All participants were recruited using the online recruitment service *Prolific*,[2] and they were paid £6.00 per hour, the recommended minimum pay. They were informed about the voluntary nature of their participation, having the freedom to withdraw at any point, and of their rights including the right to access, rectify, and delete their information. They were also shown the information sheet describing the research objectives, methodology, risks, and benefits. Informed consent was obtained from all participants.

3.2 Prescreening

The prescreening of the participants was performed as a two-step process: first, using predetermined criteria that are available in the recruiting platform (Prolific), and then, based on a questionnaire.

In the first step, we selected participants based on age (18–42), nationality and country of residence (Italy, Spain, and Portugal), fluent in English, who had participated in at least 20 surveys in Prolific, and with a task approval rate above 90%. In terms of gender, sex, and education level, no filter was applied. We chose to limit age including only Millennials and Generation Z subjects, i.e., those born between 1981 and 2012, known to have a predilection for EM [57, 68, 95], but also to narrow the generational differences among participants. The selection of only three countries in Southern Europe was motivated by the idea of having participants: (1) with a relatively similar cultural background; (2) living in the same time zone (GMT +1/+2). This last factor was fundamental to facilitating the daily interaction between participants and the researchers, the authors live in a country within the same time zone as the study participants.

Subjects matching the aforementioned criteria were redirected to the second part of the prescreening that consisted of a questionnaire in *PsyToolkit* [86, 87], a web-based framework to conduct psychological surveys and experiments. The questionnaire was composed of three main parts. In the first part, we asked participants to optionally confirm their demographic information. This step was included to double-check the reliability of the information provided by Prolific. In the second part, we asked for additional information about participants' listening habits. In detail, they self-assessed with three 5-point Likert items: (1) taste variety, (2) EM listening frequency, and (3) EM taste variety. Additionally, we asked them to indicate the preferred music streaming platform and the average daily time spent listening to music. This information was used to filter out participants who (1) self-declared to listen to EM very often, (2) who indicated listening to music less than an hour a day, and (3) who did not select Spotify as the preferred streaming service. Even if including only Spotify's users may be seen as a limitation in terms of generalizability of the results, this latter condition was necessary for collecting participants' listening logs through ListenBrainz, as explained in the next section. The former two conditions were designed to create a group of subjects who listen to music more than occasionally, but who are not frequent EM listeners.

In the third part, we included a test to verify the participants' familiarity with EM artists and genres. We replicated the test validated in a previous study [75], summarized hereafter. Participants had to specify whom from a list of mainstream EM artists was (i) known, (ii) possibly known, or (iii) unknown to them. The list was composed of 20 artists selected by *AllMusic*, an expert-curated online music database, as representatives of EM [3]. Afterwards, they had to do the same task for a list of EM genres, composed of 20 genres part of the Wikipedia page about EM [97]. The final score

[2] https://prolific.co

of each test, separately for artists and genres, was computed giving more points to participants who knew less popular artists (or genres). The rationale behind this is that knowing a popular EM artist or genre (such as Skrillex or *dubstep*) makes you less a connoisseur of EM at large, in comparison to knowing a less popular EM artist or genre (such as Autechre or *IDM*). The popularity of each artist and genre was computed using several signals from *Spotify*, *Twitter*, *Facebook*, *Deezer*, *SoundCloud*, and *Last.fm*, aggregated using the GAP0 metric proposed in Reference [50]. The list of artists and genres and the corresponding GAP0 score is presented in Appendix B (Figures 16 and 17). Afterward, we rank separately artists and genres by their GAP0 score, assigning then an alternative score inversely proportional to their popularity ranging from 1 for the less popular to 1/20 for the most popular. Participants' familiarity score is computed as the weighted sum of the alternative scores, where participants got (i) the entire score if the artist/genre is known, (ii) half of the score if the artist/genre is maybe known, and (iii) zero if unknown. Finally, artist and genre familiarity scores are averaged, and participants who according to this test were too familiar with EM (i.e., average familiarity score > 5) were filtered out.

In summary, the following inclusion criteria were applied for the prescreening:

- **First prescreening step**
  - *Age*: 18–42
  - *Nationality & country of residence*: Italy, Spain, Portugal
  - *Gender and sex*: No restrictions
  - *Highest education level*: No restrictions
  - *Fluent languages*: English
  - *Number of Prolific previous submissions*: > 20
  - *Prolific approval rate*: 90%
- **Second prescreening step**
  - *Taste variety (self-declared)*: No restrictions
  - *EM taste variety (self-declared)*: No restrictions
  - *EM listening frequency (self-declared)*: Not very often
  - *Preferred music streaming platform*: Spotify
  - *Average daily listening time (self-declared)*: > 1 hour
  - *Average EM familiarity score*: < 5 (over 10.5)

This prescreening allowed us to select a population of listeners quite homogenous in terms of demographics, listening habits, and familiarity with EM. Indeed, our main goal is to study the impact of EM recommendations on people who are not experts nor huge fans of this genre. We also aimed at reducing the response variability caused by different cultural backgrounds. These two aspects, familiarity with EM and cultural background, have been shown to be at the root of different perceptions of diversity in music lists [75], and by controlling for those while selecting the study participants, we aimed at minimizing the influence of such confounding factors in the analysis. The prescreening survey is accessible in the Supplementary Materials.

3.3 Listening Logs Collection

After being selected for participating in the study, participants were asked to create an account on ListenBrainz, allowing the collection of their listening logs for the entire duration of the study. ListenBrainz is a platform that keeps track of what music its users listen to and provides them with insights into their listening habits. It is operated by the *MetaBrainz* foundation,[3] a non-profit organization that has been set up to build community-maintained databases and make them

[3] https://metabrainz.org

Table 1. Guttman Scale Built Using the Question "Would You Be Open to Listening to One Hour of Electronic Music" for Measuring the Openness in Listening to EM

monthly	biweekly	weekly	twice a week	every day	*o-score*
0	0	0	0	0	**0**
1	0	0	0	0	**1**
1	1	0	0	0	**2**
1	1	1	0	0	**3**
1	1	1	1	0	**4**
1	1	1	1	1	**5**

0 Represents a Negative Answer, while 1 is an Affirmative Answer.

available in the public domain or under Creative Commons licences. Data is collected complying with the GDPR, and more information about MetaBrainz's privacy policy can be found online.

Among the options for submitting the music listened to, it is possible to link ListenBrainz to the Spotify account. One of the advantages of this approach is the reliability of the metadata accessible for each log. Indeed, once a log is collected by ListenBrainz through its link with Spotify, the associated Spotify track ID and artist ID are available. With the retrieved IDs, by using the Spotify Web API,[4] it is possible to obtain several types of data, from the acoustic properties of a track to the genres associated with the artists. However, this data collection method has some drawbacks.

First, creating a ListenBrainz account is a time-consuming task for which participants need to be paid, increasing the cost of the study. Moreover, people may be reluctant to link their ListenBrainz and Spotify accounts for privacy reasons. Last, while the use of the Spotify API is quite accepted in the Music IR and RS communities, the proprietary nature of the algorithms behind the API makes it difficult to know exactly how the data is generated. Nonetheless, by using ListenBrainz, we aimed to foster the reproducibility of our study, but also to ensure the availability of the collected data for future works making them publicly available through the ListenBrainz API.

3.4 Electronic Music Feedback Questionnaire

The EMF questionnaire is designed to measure implicit and explicit attitudes towards EM. Participants completed it at the beginning, four times while being exposed to the music recommendations, and the last time at the end of the study, for a total of six times (see Figure 1). In particular, the EMF measures: (1) the participants' openness in listening to EM; (2) the valence of their implicit association with EM; (3) the stereotypes they associated with EM. It is implemented in PsyToolkit, and the time needed to complete it is approximately 10 minutes. We now continue describing separately the three parts of the questionnaire.

3.4.1 Measuring Openness. Openness in listening to EM is measured using a dichotomous Guttman scale [33]. It is a unidimensional ordinal cumulative scale for the assessment of an attribute, in this study, namely, the openness in listening to EM. In detail, subjects are asked if they would be open to listening to one hour of EM, selecting *Yes* or *No* to the following options: (a) once every month; (b) once every two weeks; (c) once a week; (d) twice a week; (e) every day. The ordinal nature of the scale suggests that a participant answering *No* to the first option naturally would answer *No* to the following questions, as shown in Table 1. The score of this scale ranges from 0 for participants declaring to be not open to listening to even one hour per month of EM, to 5 for participants affirming to be open to listening to EM one hour every day. Among the advantages of using the Guttman scale are its compact form and pretty intuitive nature while analyzing the

[4] https://developer.spotify.com

scores, apart from the ease of computing the score by simply looking for affirmative answers from the participants. Morover, while the use of other measurements, for instance, Likert-scale, would have provided only the agreement or disagreement with being open to listening EM, thanks to the Guttman scale, we have been able to identify where participants' view lies on a continuum between their openness in listening EM every day and not listening at all. Being interested in understanding how music recommendations with different levels of diversity may affect the participants' openness in listening to EM, we used the score obtained from the Guttman scale as one of the variables of the longitudinal analysis, for the rest of the article referred to as **o-score**.

3.4.2 Measuring Implicit Association. People's conscious judgment represents only a facet of how evaluative associations are experienced. This is why we included in the questionnaire the **Single Category Implicit Association Test (SC-IAT)** [46], a variant of the more famous **Implicit Association Test (IAT)** [30]. The IAT aims at measuring implicit attitudes, defined as "actions or judgments that are under the control of automatically activated evaluation, without the performer's awareness of that causation" [30]. By measuring the response latencies in a categorization task, the IAT evaluates the strengths of associations between concepts, using complementary pairs of concepts and attributes. For instance, IAT has been used to measure people's positive or negative associations with Women and Men, Black and White people, or Transgender and Cisgender people. Several examples of IATs can be found on the Project Implicit webpage.[5]

In the music field, Clarke, Vuoskoski, and DeNora [14, 96] made use of the IAT for measuring if mere exposure to music may evoke empathy towards unknown cultures. Their findings support the hypothesis that, even without any accessible semantic content, listening to music can evoke empathy and affiliation in listeners. Inspired by their results, we chose to implement an SC-IAT to understand if exposure to EM may influence the implicit association of listeners. The use of SC-IAT rather than IAT has been motivated by the absence of a complementary category to EM.

In summary, using the keyboard participants have been asked to categorize as fast as possible: (1) pleasant words (*Joy*, *Love*, *Peace*, *Wonderful*, *Pleasure*, *Glorious*, *Laughter*, *Happy*); (2) unpleasant words (*Agony*, *Terrible*, *Horrible*, *Nasty*, *Evil*, *Awful*, *Failure*, *Hurt*); (3) EM genres (*Dubstep*, *Techno*, *Electronica*, *Hardcore*, *Vaporwave*, *Breakbeat*, *Electroacoustic*, *Downtempo*). By measuring the time they employed in categorizing these words correctly, we evaluated participant associations' valence towards Electronic Music. The outcome of this test is referred to as **d-score**, which takes negative values if a negative valence is associated with EM and positive values in the opposite case. Karpinski and Steinman [46] provide a detailed description of the SC-IAT design, the formula for computing the d-score, and proof of its reliability and validity. The test is accessible in the Supplementary Materials.

3.4.3 Measuring Stereotypes. The goal of this part of the EMF questionnaire is to measure what listeners opine on three kinds of stereotypes: (1) the context wherein they listen to EM; (2) the musical properties they associate with EM tracks; (3) the characteristics of EM artists they think are prominent. Responses are collected by using 5-point Likert items. First, participants are asked in which contexts they would listen to EM presenting a list of eight activities (*Relaxing*, *Commuting*, *Partying*, *Running*, *Shopping*, *Sleeping*, *Studying*, *Working*), selecting an option ranging from *Totally Disagree* to *Totally Agree*. To analyze the stereotypes associated with EM tracks' musical properties, we ask questions about: (a) *tempo* (0: mostly slow, 5: mostly fast), (b) level of *danceability*, (c) presence of *acoustic instruments*, e.g., violin, trumpet, acoustic guitar, and (d) presence of *singing voice parts* (0: mostly low, 5: mostly high). The reason why we selected these four features is twofold. First, they exemplify some of the stereotypes usually associated with EM, e.g., it has a

[5] https://www.projectimplicit.net

Fig. 2. Flow diagram representing a listening session.

fast tempo, high danceability, and low acousticness. Second, these features are among the ones retrievable at a track level (see Appendix A).

Last, participants' feedback on which characteristics they associate with Electronic Music artists is collected, focusing on: *gender* (0: mostly women or other gender minorities, 5: mostly men); *skin color* (0: mostly white-skinned, 5: mostly dark-skinned); *origin* (0: mostly low-income/developing countries, 5: mostly high-income/developed countries); and *age* (0: mostly under 40, 5: mostly over 40). Considering the nature of the study, no impact on the answers related to the artists was expected to be caused by the exposure, because no information about the artists was provided during the listening session. Nevertheless, understanding the EM artists' characteristics that participants felt to be more representative is a complementary perspective on the stereotypes they associated with EM. Even if not influenced by the provided music recommendation, these questions gave us further insights into the picture of EM that participants had in their minds.

3.5 Listening Sessions

Participants' exposure to EM recommendations took place during the 20 daily listening sessions part of the *COND* stage. Being the sessions proposed on a daily basis, their design favored the easiness and rapidness of completing the proposed task, described as follows: As an initial step, a 30-second audio clip was presented to calibrate the audio volume to avoid exposing participants to extremely loud audio potentially damaging their hearing. Afterwards, we explicitly asked participants to entirely listen to a 3 minutes audio containing a mix of a few excerpts of EM tracks, asking them to be sure to be in a quiet environment and to allow themselves to be immersed in the music.

After each listening section, the participants provided their feedback on the music listened to. Specifically, they indicated with a 5-point Likert item if they liked or disliked the music, and they selected if the listened music was familiar or not. With that, the mandatory part of the listening session ended. Following, on a voluntary basis, they were asked if they wanted to explore the playlist with the full tracks part of the audio previously listened to. If selecting *Yes*, then they were redirected to a page with a link to a YouTube playlist. If they were not interested in discovering, then the listening session ended. The time needed to complete the mandatory part of each session was approximately 5 minutes. It is important to note that the interaction with the playlists was declared to be completely optional and did not affect the participants' payment, i.e., they were not paid for the extra time spent interacting with the playlists. Figure 2 depicts the structure of a listening session.

This design allowed us to collect several types of data. First, the explicit liking and the familiarity ratings of the tracks listened to. Second, the willingness of discovering more about such tracks by choosing to explore the playlists. Third, by checking the YouTube playlist views, we also had a further metric of the actual interaction with the music proposed daily. While the definition of "legitimate view" in YouTube is not transparent [101], a common belief is that, to increase views, a user has to click the play button to begin the video, and the video has to be played for at least 30 seconds. Even if we cannot validate these hypotheses, we double-checked that simply accessing a YouTube playlist without listening to any tracks does not increase the view count of such playlists.

We chose to avoid redirecting participants to Spotify playlists to not confound the listening log collections and the platform usage. Indeed, Spotify could have registered the signal of participants' interaction with such playlists influencing eventually future recommendations by including EM related to the study. Last, YouTube playlists were not made public but accessible only to the study participants to avoid affecting the view count with interactions from external users.

3.6 End-of-study Survey

Over the course of the study, we collected several types of feedback, implicit and explicit, quantitative and qualitative, that we used to understand the impact of music recommendation diversity. As the last step, we included a final survey to collect participants' overall feedback on their participation.

The survey is composed of four sections. The first section (16 Likert items) includes questions about the participants' relationship with EM *before* participating in the study. The second (16 items) collects feedback about the experience of participants *during* the study. In the third (21 items), we ask about participants' feelings about EM *after* the study. Each of these sections is built to investigate the participants' openness, appreciation, and willingness to discover EM. Besides, a few questions are related to stereotypes associated with this genre (e.g., EM has a fast tempo or it is mostly for partying), while others are about the perceived variety of the genre itself. Last, we insert a fourth section (8 items) to understand the overall impression of participating in the study. To avoid acquiescence bias, we balanced the number of positive and negative statements in each section, and items were presented in a randomized fashion to avoid order effect bias. The time needed to complete this survey is approximately 10 minutes. The complete list of items is presented in the Supplementary Materials.

We acknowledge that, as the questionnaire took place at the end of the study, participants may have not been able to exactly recall their relationship with EM before the study, and this could limit the reliability of quantitative measurements. Nevertheless, we include it to make participants reflect on their experience in participating in the study and on how such experience could have modified their vision of EM. Under this lens, we use this survey as a tool to gain a better understanding of the validity of the study and to improve future studies' design.

4 DIVERSITY-AWARE MUSIC RECOMMENDATIONS

This section outlines the semi-automatic procedure to create the diversified recommendations to which participants were exposed during the listening sessions. In Section 4.1, we report a summary of the several steps carried out to gather the material for creating the recommendation, including in Appendix A a detailed description of the dataset and the audio models. Instead, Section 4.2 describes the designed diversification strategy.

4.1 Material

The first step was to create a dataset of candidate EM tracks to be included in the listening sessions. The goal was to select a set of tracks covering as many EM styles as possible to create a varied representation of the EM culture, however, without the presumption of including every existing nuance of it. We consulted *Wikipedia* [97] and ***Every Noise at Once***[6] **(ENaO)** to select 20 EM genres with associated 165 subgenres, listed in Table 7 (Appendix B). For each of the subgenres, we retrieved from ENaO a playlist of around 100 representative tracks for a total of around 16k tracks. Then, we filtered out tracks too popular by using the Spotify popularity indicator and the YouTube view count to avoid that familiarity with the tracks could have an effect on participants' ratings.

[6]https://everynoise.com

Last, we randomly select 10 tracks (when available, if not less) for each subgenre, remaining with a final set of 1444 candidate tracks.

As a second step, we made use of **Music Information Retrieval (MIR)** techniques to (1) extract audio embedding from the candidate tracks using state-of-the-art Deep Learning models; (2) validate these models by using standard MIR hand-crafted features. Tracks' audio embeddings were extracted using EfficientNet [90] trained with a dataset of tracks annotated with Discogs[7] metadata. Among the models tested, this showed the best performance in terms of clustering the candidate tracks coherently with respect to the considered taxonomy of EM genres. Figure 13 (Appendix B) displays a 2-dimensional representation of the embedded space obtained with this model. Then, focusing on four features (*tempo*, *danceability*, *acousticness,* and *instrumentalness*), we investigated the consistency of the tracks' embedded space. In particular, we centered our attention on if track embeddings clustered together displayed similarity also in terms of the aforementioned features.

4.2 Diversification Strategy

The diversification process for creating the recommendation lists to which participants were exposed during the study was based on two main criteria. First, we aimed at controlling the *inter-list diversity* to present to one group of participants a varied selection of EM throughout the 20 listening sessions, while to the other group only a tiny fraction of EM. Second, we limit the *intra-list diversity* for both groups to avoid creating listening sessions too varied and fragmented to become potentially annoying. Having study participants not familiar with EM, to ask them to listen in sequence, e.g., to a glitchy *IDM* track and then a soft *Minimal Techno* track, could have negatively impacted their listening experience, consequently affecting the drop-out rate.

Other design choices of the recommendations were the following: First, each list had to be formed by four tracks to limit the length of the listening session. Second, each list had to contain tracks with no more than three different genres. Indeed, even if the tracks' embedding preserves some musical characteristics of EM genres (see Appendix A), it is also true that two tracks could be very near in the embedded space even if labelled differently.

Having in mind these aspects, we implemented the following strategies: To minimize the intra-list diversity, we found for every track in the dataset the three nearest tracks, according to the pairwise cosine distance between embeddings. These quadruples of tracks were the candidate recommendation lists. Afterwards, to minimize the inter-list diversity, we selected 20 quadruples from a single genre (Figure 3(A)), while, to maximize it, we selected one quadruple for each of the 20 genres in the dataset (Figure 3(B)).

We selected *Trance* as a seed genre of recommendation lists with low inter-list diversity (from now on, simply **low diversity** or **LD**) for the following reasons: First, it has a number of subgenres that ensure some variability between lists, without exceeding like in the case of House (see Table 7). With regard to this aspect, further valid choices could have been *Techno*, *Ambient,* or *Hardcore*. However, the *Trance* tracks' cluster has, on average, a higher silhouette score (Figure 12), meaning that in the embedded space, tracks were better clustered than the other genres. This aspect was important to ensure to have enough candidates for creating coherent listening sessions with low intra-list diversity. Second, *Trance* music reflects some stereotypes of electronic music: quite danceable, with almost no use of acoustic instruments and a small presence of vocal parts. Hence, participants exposed to LD recommendations interacted with a stereotypical idea of the EM culture, with few variations in terms of musical properties. Even if differences between sessions existed because of different characteristics of the *Trance* subgenres, a certain level of homogeneity

[7] https://discogs.com

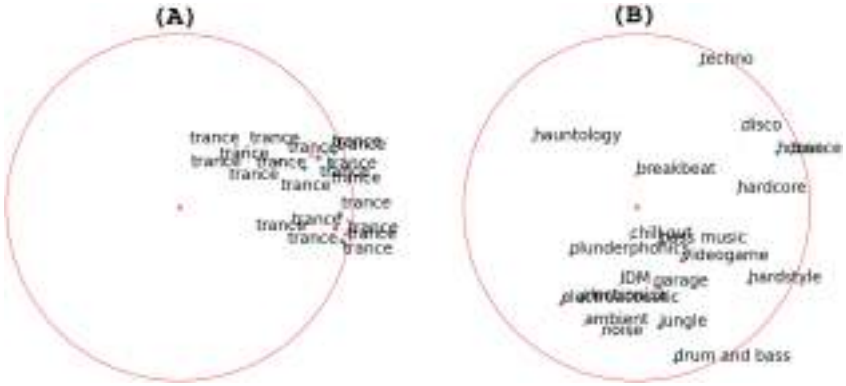

Fig. 3. Diversification outcomes represented in the 2-dimensional embedded space. Each point represents the average position of a list's tracks. (A) displays the lists with low inter-list diversity (LD), while (B) lists with high inter-list diversity (HD).

was ensured by having selected a single genre to create the recommendations. It is important to note that the use of a genre different from *Trance* could have affected the results, especially in terms of participation and engagement in the study. Indeed, *Trance* music reflects few stereotypes associated with EM, meaning that its sounds are somehow expected also for people not familiar with this genre. We believe that using a more sophisticated and less popular genre, e.g., ***Intelligent Dance Music (IDM)***, could have made it more difficult for the participants to go through 20 listening sessions, consequently increasing the drop-out rate.

On the contrary, to obtain high inter-list diversity (from now on, simply **high diversity** or **HD**), we picked a quadruple of tracks for each different genre, which was enough to ensure that participants exposed to such recommendations would explore different facets of EM. These two diversification strategies effectively led to statistically significant differences between recommendation lists. In detail, we tested these design objectives:

— The average *inter-list diversity* should be higher for the HD recommendations in comparison to the LD recommendations.
— The average *intra-list diversity* should be similar for each recommendation created using the two diversification strategies.
— The average difference between median values of the hand-crafted features should be higher in the HD recommendations than in the LD ones.

According to the statistics presented in Table 2, the aforementioned design objectives were satisfied. The objectives are further validated by the boxplot presented in Figure 18 (Appendix B), which shows the features' distribution.

Thanks to the described procedure, we created recommendation lists formed by 4 tracks, 20 with high and 20 with low diversity. In every list, tracks were selected to minimize their differences according to the distance between their embeddings. The final step was to mix the four tracks in each recommendation list to create a single audio file to be included in a listening session. To accomplish that, we implemented the following procedure using *Pysox*, a Python wrapper around Sox [8]: First, we randomly selected a 45-second excerpt for each track in the list. The sample rate of each excerpt was converted to 48,000 Hz and the volume normalized to −3db. Then, we joined the excerpts together, including a 1-second fade-in and fade-out, to help the listeners recognize the start and the end of each track in the mix. At the end of this process, we obtained a 3-minute mp3 audio for every recommendation list.

Table 2. T-test Results Comparing High Diversity (HD) and Low Diversity (LD) Recommendation Lists

	T-stat	df	p-val	CI 95%	Cohen-d	power
Inter-list diversity	22.52	378	<0.01	[0.27, 0.33]	2.31	1.00
Intra-list diversity	2.25	38	0.03	[0.0, 0.07]	0.71	0.59
Tempo	9.92	378	<0.01	[7.83, 11.7]	1.02	1.00
Danceability	10.56	378	<0.01	[0.16, 0.23]	1.08	1.00
Acousticness	9.99	378	<0.01	[0.06, 0.09]	1.02	1.00
Instrumentalness	6.05	378	<0.01	[0.12, 0.23]	0.62	1.00

P-values are corrected by using the Holm–Bonferroni method.

A final manual check of each audio was done to ensure that each listening session was properly mixed. We additionally verified that the content in each audio was appropriate for the purpose of the study. Indeed, it is not uncommon that lyrics in EM could contain references to sex, use of drugs, or blasphemy. While we do not want to advocate for a moral judgment of the artists' forms of expression, still we agreed that it was necessary to remove some tracks to respect every subjectivity that might eventually participate in the study.

5 RESULTS

This section introduces the results of the analysis of participants' feedback and listening logs collected during the 12 weeks of the experiment. We start by describing in Section 5.1 the population of our study, commenting on demographics, listening habits, and familiarity with EM, including in Section 5.1.1 the analysis of the data retrieved from ListenBrainz. Then, we continue in Section 5.2 reporting on the participants' group assignments, participation, and drop-out rate. In Section 5.3, we analyze participants' feedback during the 20 listening sessions. Afterwards, Section 5.4 includes the longitudinal analysis of the EMF questionnaire's responses, focusing first on openness and implicit association, and then on EM stereotypes. Last, Section 5.5 presents the results of the End-of-study survey.

5.1 Participants' Demographics and Listening Habits

The exploratory nature of the study, and the lack of any meta-analysis that defines the ground truths of our variables, made it difficult to estimate with a power analysis the exact number of participants needed to observe potentially existing statistical differences. Nonetheless, guidelines from Human-Computer Interaction research helped us in estimating a valid number of participants [11, 15, 45]. Indeed, when performing a t-test (or an equivalent non-parametric statistical hypothesis test) for the difference between two independent means, to observe a medium effect size (an effect likely to be visible to the naked eye of a careful observer), at a significance level of $\alpha = .10$ (acceptable for an exploratory study), with a statistical power of .80 (to avoid incurring a too-great risk of a Type II error), according to Cohen [15], a sample of size 100 participants is required, i.e., 50 participants for each group. This is the reason why, following the prescreening, we recruited **110 participants** for this study, also foreseeing the eventuality of some of them dropping out.

Most of the selected participants are aged between 18 and 32 years (91%), come from Portugal (61%), Italy (32%), or Spain (7%), almost equally divided into binary genders (53% women and 44% men)[8] with a small fraction of non-binary participants (3%). In terms of education, 69% have a bachelor's degree or lower, and 63% of the participants are still enrolled on a study course (bachelor, master, etc.). According to their self-declared answers, 88% affirm having varied listening habits,

[8]Women and men include both cisgender and transgender identities.

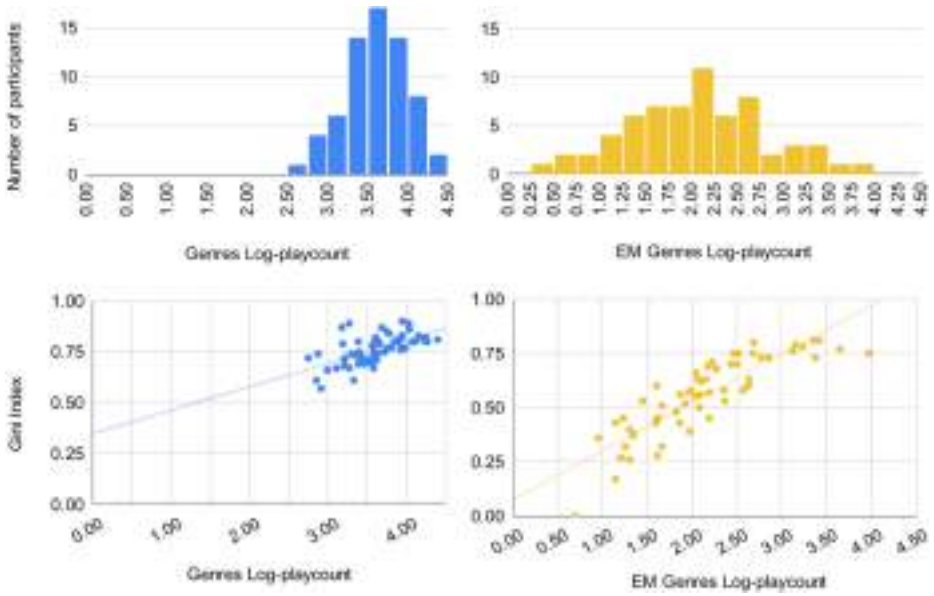

Fig. 4. Distribution of the participants' genres log-playcount (*top*) and genres log-playcount versus Gini index (*bottom*), computed with the whole set of logs (*left*) and only with EM logs (*right*).

only a quarter affirms listening often to EM, and a third affirms listening to a varied selection of EM. In terms of listening time, 78% declare to daily listening, on average, between 1 and 3 hours of music.

The two familiarity tests (artist and genre; see Section 3) help us in estimating the participants' knowledge of the EM scene. The test score ranges from 0, if no item in the lists is known, to 10.5 if all items are known. In the artist test, the average score obtained by the selected participants ($N = 110$) is 1.7 ± 1.2. Instead, the average score of all the participants in the prescreening ($N = 437$) is 2.6 ± 2.1. In terms of genre familiarity, the average score is 5.1 ± 1.6 against 5.7 ± 1.9. Averaging over the two tests, the recruited participants got 3.4 ± 1.1 against 4.2 ± 1.8. As these numbers evidence, the filter applied during the prescreening made it possible to select a group of participants, on average, less familiar with mainstream EM artists and genres in comparison to a wider group of listeners with shared demographics. In the next section, we further characterize the study participants by analyzing their listening logs.

5.1.1 Listening Logs Analysis. The ListenBrainz listening logs give us an alternative perspective for understanding the relationship that the participants have with EM. Unfortunately, the analysis comprehends only data from 66 accounts, because technical issues limited the stability of the connection between Spotify and ListenBrainz, which eventually made it impossible to collect data from all the participants. Nonetheless, the collected data are representative of some trends that we describe hereafter.

We start by analyzing the participants' log-playcount over the course of the study, considering first the whole set of listening logs, and then only the EM logs.[9] On average, participants listened to 1,479 tracks over the course of the study, more or less a daily hour of music if considering 3-minute long tracks. In contrast, during the study they listened, on average, only to 56 EM tracks,

[9]We consider as *EM log* a track listened to by a participant that is composed or performed by an artist associated with one or more EM genres among the ones retrieved from the Spotify API.

meaning more or less 1 EM track a day. Figure 4 (*top*) displays the distribution of the genres log-playcount separately over the whole tracks (*left*), and only for EM tracks (*right*). We may observe that several participants have an EM log-playcount lower than 1, i.e., they listened to less than 10 tracks associated with EM over the course of the study. Overall, for more than half of the study participants EM represents less than 15% of the whole music listened to.

Moreover, to understand the variety of music genres they listened to, we compute the Gini index separately with the whole set of listening logs, and only with the EM logs. For the latter set, the average Gini index (0.43 ± 0.20) is smaller than the one computed with the whole set of logs (0.62 ± 0.09). This indicates that participants seem to have, on average, more varied habits in terms of EM in comparison to the whole music they listened to. Figure 4 (*bottom*) shows the relationship between the Gini index and log-playcount. We observe that they are positively correlated ($\rho = .62, p < .01$), with an even stronger linear correlation between the two variables in the case of EM logs ($\rho = .86, p < .01$). This supports the idea that the more a participant listened to EM, the less varied it was. Hence, also taking into account the self-declared and estimated not-expertise of participants with respect to EM, we hypothesize the presence of two groups of listeners in our study: (a) occasional heterogenous EM listeners, with a low Gini index (0.0–0.5) and low log-playcount (0–2); (b) more-than-occasional homogenous EM listeners, with a high Gini index (0.5–1.0) and high log-playcount (2–4).

What is shown until now gives us an idea of the participants' listening habits, which can be further analyzed by looking at the kind of Electronic Music they mostly listened to during the study. Figures 19, 20 in Appendix B display the top genres and artists ranked by popularity in the participants' logs, from which we may infer some trends. Indeed, we note that *House* and *EDM* alone constitute 75% of the EM that participants listened to over the 12 weeks. Such results do not come as a surprise, because, as previously commented, the demographic segmentation of the participants in our study to some extent is similar to the one of *EDM* listeners [95]. The most frequent artists in the logs, *David Guetta*, *Avicii*, and *Alok*, among the most popular in the scene, confirm the preference of the participants for mainstream EM.

Based on these observations, we may draw the following picture of the population of our study: They are Millennials and Gen Z from Southern Europe, equally divided into binary genders with a small fraction of non-binary, mostly without a graduate degree and still studying. Average in terms of time spent listening to music and having a self-appearance of having heterogeneous preferences, they affirm to not be heavy listeners of EM nor particularly varied in terms of EM listened to. Familiar with the mainstream EM artists and genres, but far to be considered experts of the genre, they may be grouped into occasional heterogenous EM listeners and more-than-occasional homogenous EM listeners, mostly listening to *House* and *EDM*.

5.2 Grouping, Participation, and Dropout Rate

After selecting the participants matching our prescreening criteria, we needed to split them into two groups, one to be exposed to recommendations with high diversity and one with low diversity. During the pilot study, we realized that, given the size of the sample, randomly assigning participants to groups could result in an unbalanced baseline for the variables we were interested in studying. For instance, one group could have been formed by participants much more open to listening to EM than the other group. To avoid imbalance between characteristics, which may have affected the impact of the recommendations, we assigned participants using *covariate adaptive randomization* [88], creating two groups balanced in terms of familiarity with EM (*familiarity test score*), openness in listening to EM (*o-score*), implicit association with EM (*d-score*), and the number of EM tracks listened to during the *PRE* stage. At the end of this process, we got a group of 55 participants assigned to the HD group and a group of 55 participants assigned to the LD group.

Table 3. Summary of the Participation in the Study

	PRE	COND1	COND2	COND3	COND4	POST
Electronic Music Feedback (EMF)						
HD	55 (100%)	47 (85%)	46 (83%)	44 (80%)	46 (83%)	44 (80%)
LD	55 (100%)	50 (90%)	49 (89%)	47 (85%)	50 (90%)	46 (83%)
All	**110 (100%)**	**97 (88%)**	**95 (86%)**	**91 (82%)**	**96 (87%)**	**90 (81%)**
Listening Sessions (LSs)						
HD	—	50 (90%)	47 (85%)	48 (87%)	48 (87%)	—
LD	—	52 (95%)	50 (90%)	52 (95%)	49 (89%)	—
All	—	**102 (92%)**	**97 (88%)**	**100 (90%)**	**96 (88%)**	—

LS values are the median participation over each week.

Table 3 summarizes the participation of the two groups in the EMF questionnaire and in the listening sessions. Not surprisingly, participation decreased over the course of the study, notably between the *PRE* and *POST* phase in the EMF questionnaire (−19%) and less between the first and fourth week of the listening sessions (−4%). Overall, we may notice a small difference between the two groups, with the LD participants being more active during the study than their HD counterparts. Some participants have been excluded during the course of the study if (a) they never showed up after being selected (*initial nonresponse*) or (b) after a certain point they stop participating (*attrition*). The initial nonresponse was quite high for the EMF questionnaire (HD: 7%, LD: 5%), but relatively small for the listening sessions (HD: 5%, LD: 0%). Instead, attrition was equal for both tasks (HD: 7%, LD: 2%). These numbers are in line with retention rates observed in Prolific and generally in longitudinal studies [17, 49].

In conclusion, we excluded from the analysis the responses of the participants who did not participate in: (a) more than 4 listening sessions and (b) more than 3 EMF questionnaires. With this choice, we ended up analyzing the responses of 94 over 110 participants (85%), 45 over 55 in the HD group (82%), and 49 over 55 in the LD group (89%).

5.3 Listening Sessions Analysis

During the 20 listening sessions in the *COND* stage, we collected four types of data to measure the impact that recommendations had on the participants of the HD and LD groups:

— *Playlist accesses*: assigning 1 to a participant who after a session chose to explore the session's playlist, 0 otherwise.
— *Playlist interactions*: YouTube playlist's view count, aggregated over each group of participants.
— *Like ratings*: ranging from −2 if a listening session was totally disliked by a participant to 2 if a session was totally liked.
— *Familiarity ratings*: 1 if the tracks in a listening session were familiar to a participant, −1 if unfamiliar, 0 if unsure.

Figure 5 displays the distribution of playlists' accesses and interactions, where trend lines are computed with the linear least squares method. Similarly to the overall decrease in study participation, we may note that over the course of the sessions the overall engagement with the playlists decreased. Looking at the number of accesses and interactions, the HD participants seem to be more interested in discovering the music listened to than the LD ones. Moreover, we notice that on several sessions LD participants accessed a playlist but had zero interactions with it (e.g., sessions 10 and 11). This phenomenon never occurred to HD participants, who, on the contrary, in some sessions had much more interactions than accesses (e.g., sessions 5 and 17), meaning

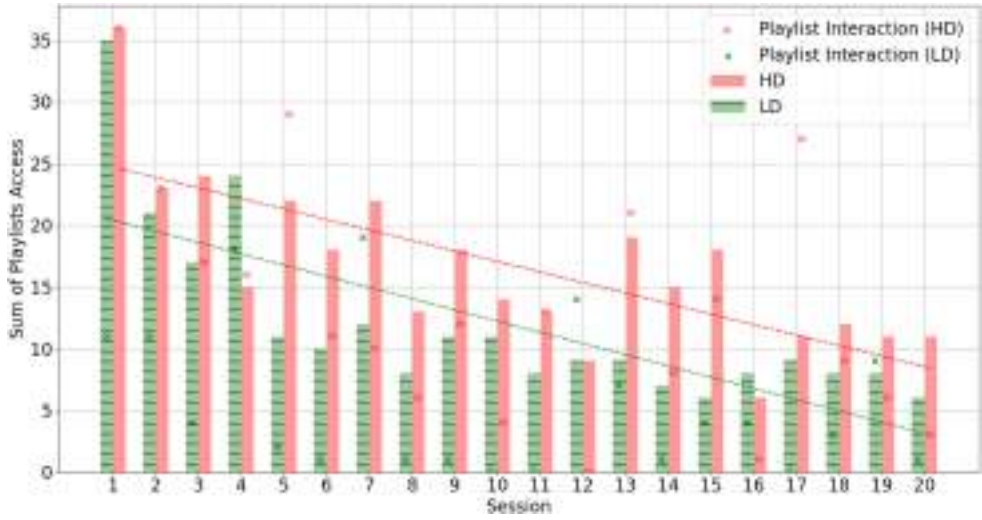

Fig. 5. Distribution of the playlists' accesses and interactions.

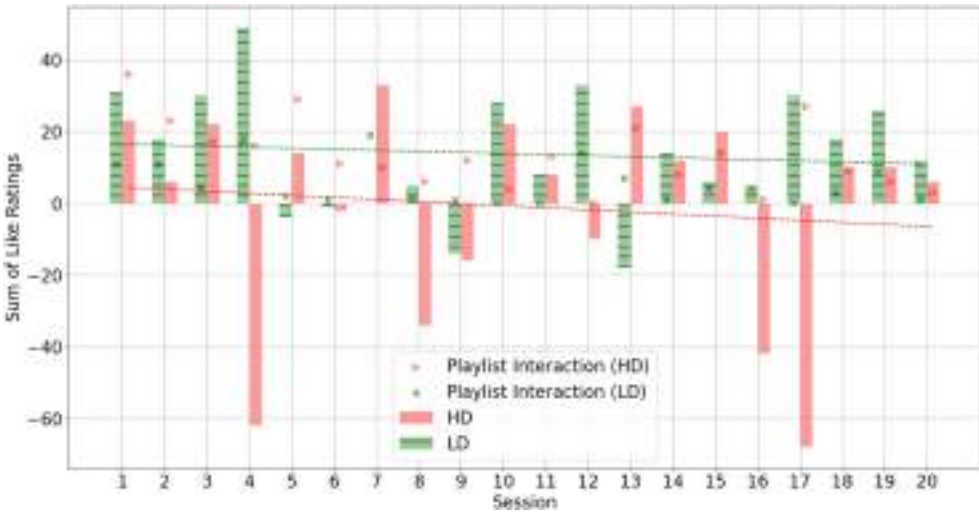

Fig. 6. Distribution of like ratings and playlists' interactions.

that some participants interacted several times with the same playlist. Naturally, in both groups, accesses and interactions are positively correlated (HD: $\rho = .75$, LD: $\rho = .51$).

Figures 6 and 7 merge the playlists' data and the like ratings, giving us an alternative view for understanding the participants' reception of the recommendations. In Figure 6, it emerges that the HD group disliked more sessions and with more extreme ratings compared to the LD group, which, on average, seem to have appreciated most of the music they have been exposed to. Nevertheless, HD participants, even when they mostly disliked a session, interact with the playlists on YouTube (e.g., sessions 4 and 17).

Such behavior is confirmed in Figure 7, where like ratings are split between participants who accessed the playlists (*light bars*) and those who did not (*dark bars*). We see that some of the HD participants choose to access the playlist and discover more about the tracks listened to even in

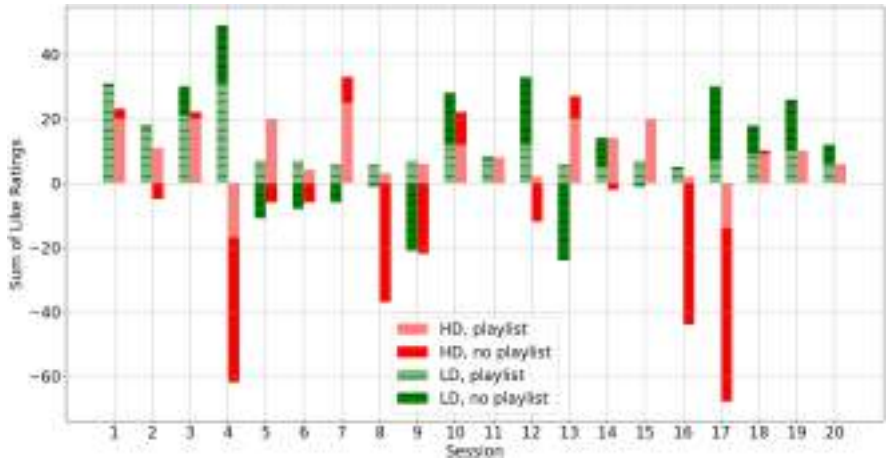

Fig. 7. Distribution of like ratings split among participants who accessed the playlist (*light bar*) and those who did not (*dark bar*).

the most disliked sessions (*negative light bars*), a behavior not present for the LD group. On the contrary, some of the LD participants, even if liking the tracks listened to, choose to not interact with the playlists (*positive bars with zero interactions*, e.g., Figure 6, sessions 10 and 11) or not access the playlists (*positive dark bars*, e.g., Figure 7, sessions 4 and 10).

The familiarity ratings' results are shown in Appendix B (Figures 21 and 22) and summarized as follows: The sum of ratings is negative for almost every session, indicating that the tracks were mostly unfamiliar to the participants.[10] This trend was expected because of participants' non-familiarity with EM, but also because of the popularity filter applied before creating the recommendations. Besides, we observe a positive correlation between like and familiarity ratings (HD: $\rho = .55$, LD: $\rho = .72$), confirming what previous music psychology scholars have extensively proven: The more familiar a track sounds, the more is likely to be appreciated.

We further confirm the previous findings by performing Mann-Whitney U tests comparing the playlist accesses and interactions, and the like and familiarity ratings of the HD and LD groups' participants, aggregated by listening session. This test is commonly used to compare the differences between two independent samples when the sample distributions are not normally distributed and the sample sizes are small (in this case, the sample size is the number of listening sessions $n = 20$). In particular, we verify that:

(1) The HD group had more accesses to playlists than the LD group.
(2) The HD group had more interactions with playlists than the LD group.
(3) The LD group liked the tracks more than the HD group.
(4) The LD group liked more tracks, even if not accessing the playlists, than the HD group.
(5) The HD and LD groups had the same level of familiarity with the listened tracks.

Table 4 reports the outcomes of the tests. **(1)** and **(2)** are confirmed by a significant difference between HD and LD groups in terms of playlist accesses and interactions. In terms of like ratings **(3)**, we see a smaller effect size with only 62% of sessions having HD group ratings lower than the LD ones. This proportion increases to 70% if looking only at the ratings of participants who did

[10]There is one exception in session 4 for the LD group, presenting a positive peak not in line with the rest of the sessions. This is due to one of the tracks included in that session "Around the World (La La La La) (Ultra Flirt Hands Up Remix Edit)," a remix of the famous track by A Touch of Class (ATC).

Table 4. Summary of the Mann-Whitney U Test Results

	Data	M_{HD}	M_{LD}	U	Altern.	p-val	CLES
(1)	Playlist access	15.0	9.0	307	greater	<.01	.77
(2)	Playlist interaction	11.5	3.5	301	greater	.01	.75
(3)	Like ratings	9.0	13.0	152	less	.20	.62
(4)	Like ratings and playlist access	−1.0	1.0	120	less	.05	.70
(5)	Familiarity ratings	−21.5	−24.0	180	two-sided	.60	.45

M stands for median value. CLES (Common Language Effect Size) is the proportion of pairs where x is higher than y. When the alternative is "greater," x are the values of the HD group and y of the LD group. With the value "less," we have the opposite scenario. P-values are corrected by using the Holm–Bonferroni method.

not access the playlists **(4)**. This confirms that even when LD participants liked the tracks in the listening sessions, they interacted with the playlists less than the HD group. Last, looking at the familiarity ratings **(5)** no significant difference is found, confirming that the design of the listening session was effective, exposing subjects to music they were mostly unfamiliar with.

5.4 EMF Questionnaire Analysis

Through the analysis of the listening sessions, we have shown the distinct reactions of the two groups of participants when receiving the music recommendations. Hereinafter, we continue by focusing on the impact of the exposure to EM recommendations, first, in terms of openness in listening and implicit association, and second, in terms of stereotypes that participants associated with this music genre.

5.4.1 D-score and O-score. The d-score measures the implicit association with EM, having negative values if a negative association is present and positive values in the opposite case. The o-score measures the openness in listening to EM, ranging from 0 if a participant is not open to 5 if a participant is extremely open. We collected these scores six times during the longitudinal study, first at the beginning (*PRE*), four times during the conditioning phase (*COND* 1–4), and last at the end of the study after 12 weeks from the start (*POST*).

Figure 8 shows the average scores and standard deviations separately for the two groups. In terms of d-score, we may observe that, starting from a slightly positive average, HD participants' score decreases toward zero. Even if with more fluctuations, similarly, the LD participants end up with an average score near zero, almost equal to the initial one. Instead, for the o-score, both groups present a slight increase comparing the *PRE* and *POST* averages, with LD presenting a higher response variance.

Table 5 reports the percentages of participants grouped by their scores. For the o-score, we split the participants into two groups, the less open in listening to EM having a score between 0 and 2 and the more open having between 3 and 5. At the beginning (*PRE*), the proportion for both HD and LD groups is around 75–25 (more open–less open), while in the *POST* measurement, the proportion of open participants slightly increases, resulting in approximately 80–20.[11]

In terms of the d-score, the proportions in the two groups are initially quite different, with the HD group having more positive scores than the LD group.[12] Notwithstanding, over the course of the study, the participants' scores move towards zero, with the neutral group (d-score

[11]Given the voluntary basis for participation, the skewness of the initial openness' distribution may be motivated by the fact that people not open to listening to Electronic Music, even if paid, may be more reluctant to participate in a longitudinal study wherein they have to listen to such music every day for one month.

[12]Even if the d-score was included as a covariate while assigning participants to groups using covariate adaptive randomization, the higher initial nonresponse in the HD group caused such imbalance.

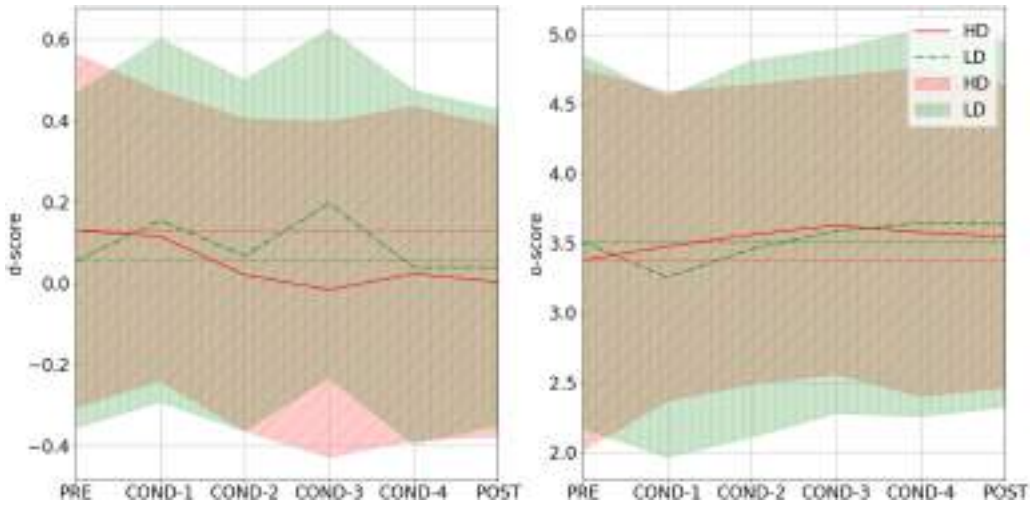

Fig. 8. Average and standard deviation of d-score (implicit association) (*left*) and o-score (openness) (*right*). The dotted lines are the average baseline measurements (*PRE*). The filled area is the standard deviation from the mean value.

Table 5. Percentage of Participants Divided According to the Scores Collected

		HD			**LD**		
		PRE	POST	*diff*	PRE	POST	*diff*
o-score	0–2	26.7%	14.3%	−12.4%	24.5%	18.2%	−6.3%
(openness)	3–5	73.3%	85.7%	+12.4%	75.5%	81.9%	+6.3%
d-score	<−0.25	17.8%	28.6%	+10.8%	28.6%	22.7%	−5.9%
(implicit	±0.25	42.2%	45.2%	+3.0%	38.8%	52.3%	+13.5%
association)	>0.25	40.0%	26.2%	−13.8%	32.6%	25.0%	−7.6%

$\in [-0.25, 0.25]$) consisting of almost half of the participants for both HD and LD groups, while the positive (d-score > 0.25) and negative (d-score < 0.25) scores are equally split.

This analysis shows few aspects of the average behavior of the HD and LD groups, without, however, considering individual differences. To further confirm what found at the group level, we explore the association between the rate of change and the initial scores by using the individual slopes obtained from the regression analysis of each participant's scores. Figure 9 shows the slopes describing the trajectory of each participant versus the baseline scores obtained at the beginning of the study, separately for the d-score and o-score. Every point in the scatter plot is computed using a single participant data, which consists of 6 d-scores (left plot) and 6 o-scores (right plot) (see lines 758-760). If a point in the left (right) plot has the coordinate of the y-axis greater than zero it means that according to the 6 measurements, the participant's d-scores (o-scores) increased over the course of the study. If the coordinate is lower than zero we have the opposite scenario.

The slopes of the d-score are mostly clustered around zero, meaning that the implicit association towards EM did not extremely change for most of the participants. In the bottom-right quadrant, a greater presence of HD participants is visible, who represent the subjects starting with a positive attitude and then moving towards more negative ones. On the contrary, in the top-left, we see mostly LD participants, representing the opposite scenario. The average slope for the LD group is almost zero, while for the HD group is negative, confirming that, overall, participants of this latter group developed less positive implicit association during the study.

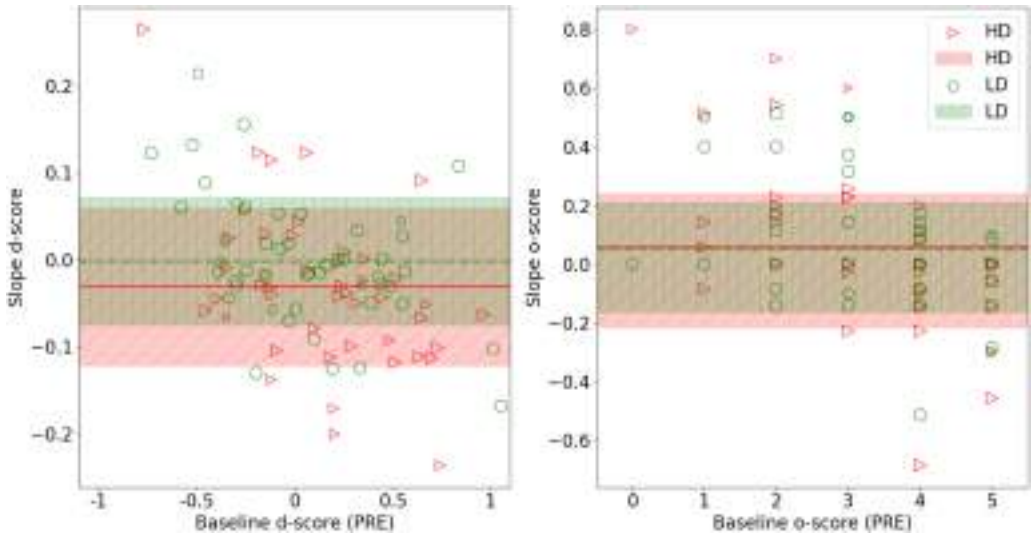

Fig. 9. Individual d-score (implicit association) (*left*) and o-score (openness) (*right*) slopes versus baseline (*PRE*) scores. The horizontal line is the mean slope for each group, while the filled area represents the standard deviation.

Table 6. Summary of the Wilcoxon Signed-rank Test Results

	Data	M_1	M_2	W	Alt.	p-val	CLES
(6)	d-score (*PRE-COND*)	.065	.008	2,519.5	greater	.02	.53
–	d-score (*PRE-POST*)	.069	.016	2,260.0	greater	.05	.55
(7)	o-score (*PRE-COND*)	4.0	4.0	414.5	less	.10	.54
–	o-score (*PRE-POST*)	4.0	4.0	434.5	less	.16	.52

M stands for median value (1: *PRE*, 2: *COND* or *POST*). CLES (Common Language Effect Size) is the proportion of pairs where x is higher than y. P-values are corrected by using the Holm–Bonferroni method.

We observe a different situation for the o-score, where no particular differences are observed between groups. Only in the case of the two participants who started the experiment declaring to be not open to listening to EM neither for one hour a month (baseline score equal to zero), we clearly observe different slopes. Indeed, the LD participant over the weeks seems to have not changed her openness; instead, the HD participant has a positive slope. This indicates that even if at the beginning she was not open to listening to EM at all, eventually she started to be more open during the study.

Starting from these observations, we further verify the impact of recommendations on the scores by using a Wilcoxon signed-rank test. In fact, over the course of the study for the whole group of participants, we observe that: **(6)** the implicit association with EM tended towards neutral valence, and **(7)** the openness in listening to EM increased. Two comparisons are made, first between the scores in the *PRE* stage and the ones at the end of the fourth week of the *COND* stage (*PRE-COND*), and then between *PRE* and *POST* stage (*PRE-POST*). Using the former, we are able to measure the impact of recommendations on participants right after being exposed to EM, while with the latter, we measure if the impact is still persistent after one month from the exposure.

Table 6 reports the outcomes of the tests. In the case of the d-score, after the exposure participants' scores tendentially decrease, a trend confirmed when looking at the differences between

the beginning and the end of the study. Instead, by analyzing the o-score, we observe an opposite behavior, having an increase right after the exposure, which becomes not significant comparing the *PRE* and *POST* measurements. However, for both scores, the effect size was not particularly large. As a further step, by means of correlation analysis, we measure the temporal stability of the two scores considering again the two intervals *PRE-COND* and *PRE-POST*. In terms of d-score, we observe lower stability over time in comparison to the o-score both in the *PRE-COND* measurements (d-score: $\rho = .30$, $p < .01$, o-score: $\rho = .57$, $p < .01$) and in the *PRE-POST* measurements (d-score: $\rho = .34$, $p < .01$, o-score: .53, $p < .01$). These results corroborate the idea that implicit measurement may be less resistant to situationally induced changes than explicit measures [28].

After highlighting the overall impact of recommendations on the study participants, we are interested in understanding the role of diversity in such change. We use two regression methods to compare the HD and LD groups. In the *follow-up analysis,* we look at the difference in the mean response at follow-up (*POST*) comparing the two groups. Instead, in the *change analysis,* we study the difference between the average change (*PRE-POST*). From the former method, we have no evidence of a significant difference in the mean responses between HD and LD groups at the *POST* stage, both for the d-score ($\beta_1 = -.04$, $SE = .08$, $p = .66$) and the o-score ($\beta_1 = -.09$, $SE = .26$, $p = .73$). Similarly, the change analysis does not evidence differences between groups in the average change between the beginning and the end of the experiment, both for the d-score ($\beta_1 = -.09$, $SE = .10$, $p = .38$) and the o-score ($\beta_1 = .1$, $SE = .27$, $p = .71$).

In summary, after the exposure to four weeks of music recommendations, we found a slight change in implicit association (**6**) and openness (**7**), but we have not evidenced any particular influence by the degree of diversity at which participants were exposed.

5.4.2 Stereotype Analysis. The results of this section of the EMF questionnaire are displayed in Figures 23, 24, and 25 (Appendix B), respectively, for the listening contexts, the musical properties, and the artists' characteristics that participants associated with EM. Hereafter, we summarize the main results. As done for the d-score and the o-score, we compare exclusively the measurements taken at the beginning of the study (*PRE*), at the end of the listening sessions (*COND*), and at the end of the study (*POST*).

Among the eight contexts presented in the survey, participants indicate that they would preferentially listen to EM while doing a dynamic and energetic activity (*partying*, *running*, *commuting*, and *shopping*). On the contrary, they disagree that EM is suitable for being listened to during activities that require a higher level of calm or concentration (*sleeping*, *studying*, and *relaxing* or *working*).[13] Nevertheless, the exposure to recommendations did not largely affect the opinion of the participants. Indeed, performing a Wilcoxon signed-rank test between *PRE-COND* and *PRE-POST*, for six contexts out of eight, no statistically significant differences ($p < 0.05$) have been found. Likewise, by using the Mann-Whitney U test, we have not found significant differences between the HD and LD participants' responses, indicating that the level of diversity did not differently affect the participants.

The only two contexts wherein we found significant differences are *running* and *shopping*. In the former case, the LD group is strongly convinced about the use of EM for running, with the percentage of agreement passing from 62% in *PRE* to 76% in *COND* and *POST*. In the HD group, we observe an opposite tendency, passing from an agreement of 73% in *PRE* to 66% and then 68% in *COND* and *POST*. These results support the idea that being exposed only to *Trance* music, a high-energy kind of music, may affect the listeners in associating EM with a high-energy kind of activity

[13]There is some ambiguity about the definition of working, which may have influenced the responses. Indeed, depending upon the type of work, it may be considered an energetic or a calm activity.

like running. On the contrary, while exploring different facets of EM, listeners may have realized that some genres are not fit for being listened to while running. Instead, in the case of *shopping*, we see that both groups start disagreeing in the *PRE* measurement (HD: 52%, LD: 62%), but then, over the course of the study, arrive at a more balanced situation between agreement, disagreement, and neutral responses. In the case of "shopping", observing responses with less extreme values makes sense because it is neither a very dynamic activity, such as "running", nor a calm activity, such as "studying".

Similarly, the musical properties that participants associate with EM tracks have not been largely affected by the recommendations. Among the four selected features, participants changed their opinion only on the presence of acoustic instruments, especially the ones in the HD group. Indeed, for them, we find a significant difference ($p = .01$, $CLES = .32$) both comparing *PRE-COND* and *PRE-POST* measurements. Moreover, the Mann-Whitney U test confirms the significant difference between the HD and the LD groups' responses both in *COND* and *POST* measurements ($p = .04$, $CLES = .61$). Observing the distribution in Figure 24, we may notice that at the beginning, 79% of HD participants disagree on the fact that EM had mostly acoustic instruments, while in the *COND* and POST measurements, only about 50% disagree. On the contrary, the percentage for the LD group remains quite stable over the course of the three months. This is consistent with the fact that the LD group has been exposed only to *Trance* music, which rarely has parts with acoustic instruments. On the contrary, HD participants listening to genres such as *Electroacoustic* may have changed their idea about the acousticness of EM.

Last, analyzing which characteristics participants associate with EM artists (Figure 25), no statistically significant differences have been found between HD and LD groups. Only in terms of age, we find a difference between *PRE-COND* and *PRE-POST* measurements.

5.5 End-of-study Survey Analysis

The analysis of the EoS survey reveals a few more qualitative insights that complement what is presented in the previous sections. First, we analyze participants' answers concerning the openness and appreciation of EM before, during, and after participating in the study. Then, we focus on a set of stereotypes to see how exposure to EM recommendations has affected the participants' opinions.

We recall that the survey is formed by three main groups of items, the first asking about the experience of participants *before* the study, the second *during*, and the third *after* the study. The survey contains 16 Likert items to measure the participants' openness in listening to EM and 16 items for measuring appreciation of EM, divided into 6 items asking for participants' beliefs before, 6 during, and 4 after participating in the study. By analyzing these three sets separately, we may get an idea about how participants perceived a change in their openness and appreciation of EM due to their participation in the study.

Figure 10 presents the distribution of the responses for the whole group of participants. We avoid reporting results separately for the HD and LD groups, because no significant difference has been found between their responses. Analyzing the internal consistency of the sets of items computing Cronbach's alpha (α), we observe an acceptable consistency in the three sets. Indeed, for the openness items, we have: *before* $\alpha = .88$, *during* $\alpha = .82$, and *after* $\alpha = .72$. For the appreciation items: *before* $\alpha = .90$, *during* $\alpha = .86$, and *after* $\alpha = .77$. Therefore, we may assume that the items properly reflect the two concepts that we aimed at measuring.

Focusing on the openness in listening to EM, we notice that participants' responses before the study were quite balanced around the neutral response, even if with high variance ($M = 2.94 \pm 1.31$). When asked about their openness during and after the study, they averagely declared to be more open in comparison with the beginning (respectively, with $M = 3.43 \pm 0.95$, and $M = 3.48 \pm$

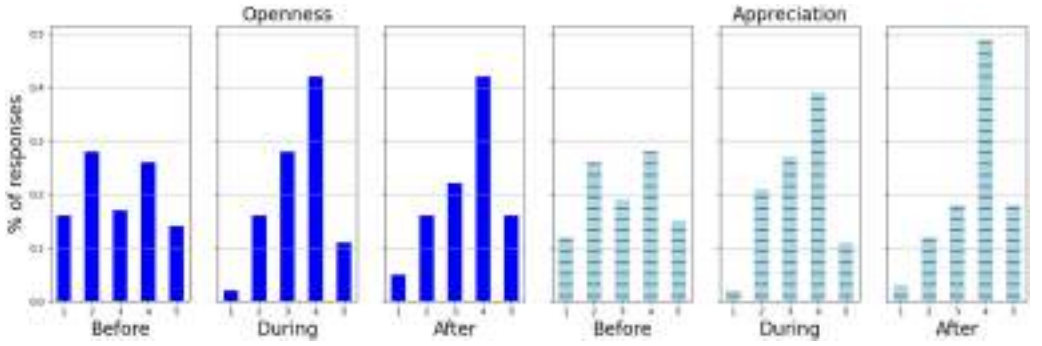

Fig. 10. Distribution of responses for the openness (*dark blue*) and appreciation (*light blue, dashed*) 5-point Likert scale. Response 1 corresponds to strong disagreement with the items, and 5 to strong agreement.

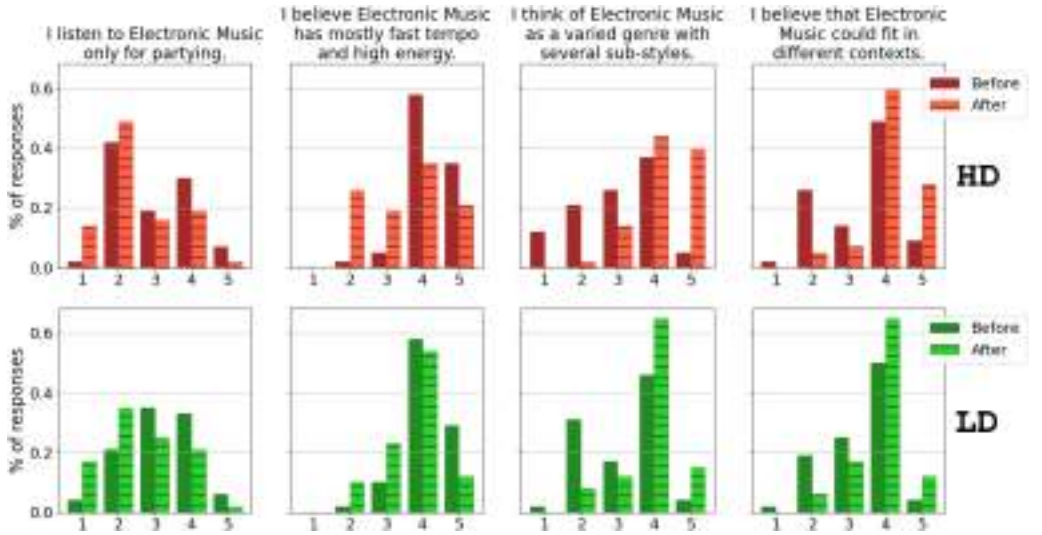

Fig. 11. Distribution of responses for the High Diversity (HD, *top*) and Low Diversity (LD, *bottom*) groups, before (*clean bar*) and after (*dashed bar*) the participation in the study. Response 1 corresponds to "Totally Disagree" with the item, and 5 to "Totally Agree."

1.08). Therefore, their perceptions are in line with the results of Section 5.4, wherein an overall increase in openness was found by looking at the results obtained from the Guttman scale. In terms of appreciation, we notice a similar trend with an initial balanced situation at the beginning of the study ($M = 3.07 \pm 1.27$) that increases towards more positive responses over the 12 weeks ($M = 3.36 \pm 0.99$ and $M = 3.67 \pm 1.01$, respectively, *during* and *after* the study).

The second focus of our analysis is on comparing four items describing four stereotypes of EM, reported in Figure 11 together with the distribution of the responses. In this case, it displays the distribution for the HD and LD groups separately, because the impact of the recommendations on these items has indeed been mediated by their diversity.

For the first item analyzed ("*I listen to EM only for partying*"), we observe that, prior to the study, participants had, on average, a balanced response towards this stereotype (HD: 2.98 ± 1.06, LD: 3.17 ± 0.97). However, during the study, the two groups realize that may be restrictive to consider EM only for partying, with an overall decrease in their average response (HD: 2.47 ± 1.03, LD: 2.56

± 1.07). In the case of the second item ("*I believe Electronic Music has mostly fast tempo and high energy*"), we observe a similar behavior. In the beginning, participants strongly agree with this statement (HD: 4.26 ± 0.66, LD: 4.15 ± 0.68), but then after the study they changed their opinion (HD: 3.51 ± 1.1, LD: 3.69 ± 0.83), agreeing less strongly to the fact that EM has a mostly fast tempo and high energy.

The next considered item ("*I think of Electronic Music as a varied genre with several sub-styles*") presents an opposite situation in comparison to the former two. Indeed, at the beginning, on average, participants neither agree nor disagree significantly with the statement (HD: 3.02 ± 1.12, LD: 3.19 ± 1.0). However, participating in the study made them change their opinion about it, agreeing much more in percentage than before (HD: 4.21 ± 0.77, LD: 3.85 ± 0.77). Similarly, also in the case of the fourth item considered ("*I believe that Electronic Music could fit in different contexts*"), we observe a neutral position of participants at the beginning (HD: 3.37 ± 1.05, LD: 3.35 ± 0.9), which later agree more with the fact that EM may fit in different contexts (HD: 4.12 ± 0.73, LD: 3.83 ± 0.72). It is interesting to note that the responses in these two latter items are the only ones for which we found a significant difference between HD and LD subjects. Indeed, by performing a Mann-Whitney U test, we obtain a p-value of .01 and .02 and a CLES of .65 and .62. This indicates that, even if for both groups the exposure to recommendation affected the beliefs about how varied EM may be and how it may fit in different contexts, in the case of the HD group, this shift has been more pronounced.

6 DISCUSSION AND LIMITATIONS

In this section, we discuss the results previously presented with the aim of answering, first, *to what extent can listeners' implicit and explicit attitudes towards an unfamiliar music genre be affected by exposure to music recommendations?* (**RQ1**); and, second, *what is the relationship between music recommendation diversity and the impact on listeners' attitudes?* (**RQ2**). We focus on four main aspects: impact on discovery (Section 6.1), implicit association (Section 6.3), openness (Section 6.2), and stereotypes (Section 6.4). Moreover, in each section, we present limitations and future work.

6.1 Impact on Discovery

The most pronounced role that recommendation diversity plays in our study is with respect to the curiosity generated in the participants when exposed to music during the listening sessions. Indeed, the listeners' willingness to explore EM playlists is significantly higher if exposed to highly diverse recommendations. Having in mind that EM was mostly unfamiliar to the subjects in our study, we believe that *when listeners are not familiar with a music genre, a diversified set of recommendations could increase their willingness in exploring such music.* A message sent to us by a study participant supports this hypothesis: "Hello! It was a pleasure participating, I discovered new artists that I really liked! Happy to have contributed to your study :) Thank you!"

Limitations. Nevertheless, the first limitation of this work rises from the definition of *discovery* itself. Indeed, if, as Nowak suggests, discoveries are "affective responses to music content that occur within individuals' life narratives and mediate their interpretation and definition of music" [69], then we argue that the responses resulting from the participation in our study cannot be compared to what listeners usually experience in less artificial situations, wherein they are exposed to music. Moreover, focusing on discovery strategies and behavioral attitudes, Garcia-Gathright and colleagues [27] show how explorative goals may vary according to the listeners' needs. Under this lens, a further step could be to link the various overt behaviors to people's affective responses to evaluate if the discovery mediated by algorithmic recommendations has some shared values with other non-digital forms of curation (e.g., a DJ tracklist created for a live event).

6.2 Impact on Openness

Throughout the study, participants' openness to listening to EM increased, indicating that, in the long term, the exposure to unfamiliar music could favor listeners in being less reluctant in approaching a genre previously unknown. This finding does not come totally anew, partly confirming results from the literature on the impact of repeated exposure to music.

Diversity here seems to motivate particularly participants who started the experiment affirming to be not open to listening to EM. Indeed, subjects who passed from not being open to being open to listening to EM for one hour or more a week are participants in the HD group who passed from not being open to being open to listening to EM for one hour or more a week are two times the ones of the LD group. Still, the impact of recommendation diversity is apparently not significant. Moreover, contrary to the implicit association, participants' openness was more consistently affected right after the conditioning stage than at the end of the study. Connecting this with the impact on discovery, we deduce that openness is highly influenced by exposure, but such influence decays rapidly when listeners stop to be exposed to recommendations. In light of this, we support the idea that *when listeners are not familiar with a music genre, repeated exposure to recommendations could increase their openness in listening to such music*. A message sent by one of the participants of the HD group right after the end of the *COND* stage supports this intuition: "Thank you so much [...] It was very interesting and a good opportunity to learn about this musical genre."

Limitations. Again, the settings of the experiment may have influenced the outcome. In fact, participants were requested to be highly focused while listening to the music during the study, a condition that is not quite frequent in today's listening practices, where music is often confined to the background while performing other activities [64]. This could have affected the openness more consistently in comparison if tracks would have been passively listened to, for instance, while walking in a mall or while working. Repeating the experiment in a non-online environment may lead to more precise results in this regard.

Furthermore, the use of the Guttman scale has its own limitations. Indeed, listeners' openness in listening to EM could be determined by an overall curiosity in discovering music. For instance, a listener with very heterogeneous musical tastes could be open to listening every day to one hour of electronic music, one hour of classical music, one hour of rock music, and more. On the contrary, a very homogeneous listener would avoid listening to electronic, classical, and rock music at the same time if disliked. We foresee the analysis of participants' openness in listening to EM with regard to their overall tendency of listening to varied music.

Nevertheless, even if the concepts of openness in listening and discovery may overlap, we believe they do represent two different attitudes. The willingness to discover unfamiliar music for the sake of curiosity does not translate automatically into being open to listening to it. In fact, discovery might eventually lead to being less open. On the contrary, someone could be open to listening to some kind of music without being willing to discover something new. In conclusion, we believe that the willingness to explore could be related to people's openness in listening, and future work should explore the relationships between these two attitudes while listening to music.

6.3 Impact on Implicit Association

The exposure to music recommendations helped participants in deconstructing part of the preexisting positive or negative association to EM, developing a more neutral attitude during the 12 weeks. The impact of diversity in this case is not significant, considering that no major differences have been found between the two groups. In fact, we argue that the tendency of the HD group to decrease more significantly in comparison to the LD group is because of the unbalance created by the different dropout rates.

Based on that, we hypothesize that *when listeners are not familiar with a music genre, to receive repeated recommendations could mitigate the valence of the implicit association with such music.* The results obtained deviate from the findings in Reference [96], where positive implicit attitudes towards facial images of people from two cultures, namely, Indian and West African, are developed by mere exposure to music from such cultures. However, the types of association and stimuli that listeners experienced in our study are undoubtedly different and presented in a different scenario.
Limitations. The following reasons may be at the root of the development of neutral attitudes towards EM in the presented study: First, and in line with the previous point on the discovery, the experimental setting mediated the affective response, also influencing the implicit association. The tendency of the participants to associate neutral valence with EM genres could be motivated by the artificiality of the events wherein participants listened to the music. Second, given the participants' unfamiliarity with the genre, it is understandable that they did not develop any positive or negative attachment to genre labels that they might have never heard of before the study.

As elegantly discussed by McLeod [61], genre namings are strictly tied to the EM communities behind which people identify, also influencing the dynamics of group formation. Therefore, we hypothesize that most of the genres shown in the SC-IAT test were only *labels* to the study participants and remained as such, given that no mechanism of bounding was incentivized. Indeed, no information about the genre of the tracks listened to was provided during or after the listening sessions. Under this lens, receiving neither positive nor negative responses from the participants may be seen as the desired result in the long term, which, however, needs further analysis to be understood in depth.

6.4 Impact on Stereotypes

What emerges from the analysis of the results is that the idea that participants have of EM reflects stereotypes usually associated with this music genre. Indeed, listeners not familiar with EM may fall into the trap of misinterpreting it as music composed only with electronic instruments (e.g., drum machine, sampler, synthesizer). Instead, acoustic instruments, even if sampled, filtered, or generally modified, have always been used by EM artists. Under this lens, the fact that acousticness is the only feature on which participants' changed their idea after the exposure to recommendations does not come as a surprise.

Moreover, the stereotype that EM is only for parties and the common belief that it has mostly a fast tempo and high energy have been partly deconstructed by the exposure to music such as *Ambient*, *Electroacoustic,* or *Chill-out.* Nevertheless, the characterization of EM as *Energetic and Rhythmic* is quite common also in scientific literature, e.g., Reference [77], and participants of the LD group at some level experienced this aspect of EM. Besides, another accomplishment of this study is to show that listeners exposed to diversified recommendations may have a better understanding of the variety of EM culture and realize that this music may fit in different contexts, not only for partying. Therefore, we deduce that *exposing listeners not familiar with a music genre to a set of diversified recommendations, showing the different facets of such genre, could deconstruct pre-existing stereotypes.* This finding is in line with the music psychology literature on the influence of music exposure.
Limitations. Likewise, the representation that participants have of EM artists is quite stereotypical, and also in this case it was something expected. EM artists are, according to them, mostly men, white, under 40, and coming from developed and high-income countries. In this case, we did not expect that recommendations would affect participants' opinions, mostly for two reasons, which however, are also two limitations of this work. First, participants' origin has been purposely restricted to a small part of the world, normally labelled as **Western, Educated, Industrialized, Rich, and Democratic (WEIRD)** societies. Similarly, also the music material part of the study

is somehow biased towards Western artists. In fact, the semi-automatic method for creating the dataset has produced recommendation lists that reflect the variety of EM as a music genre, but not its variety as music played in different regions of the world.

7 DESIGN IMPLICATIONS FOR MUSIC RECOMMENDER SYSTEMS

What has been discussed until now gives us a multifaced picture of the impact that diversified and repeated music exposure may have on listeners, suggesting some Guidance that can help practitioners in designing music recommender systems.

First, diversified recommendations are of particular relevance for enhancing discovery in streaming platforms. In offline settings, several studies have investigated the role that diversity may play [73], but to our knowledge this is the first longitudinal user study that measures the long-term impact of recommendation diversity on listeners' attitudes towards discovery. Recent works by Liang and Willemsen [55, 56] show that it is possible to favor exploration of distant music genres both in the short and long term by nudging listeners through specific design choices; for instance, presenting such genres in the top of the recommendation lists. Starting from their findings, we believe that by **mixing nudging mechanisms with diversification techniques**, practitioners may design recommendations that notably improve the experience of discovering unfamiliar music.

Second, recommender systems should sustain discovery while at the same time increase users' openness to the newly discovered music genre to support long-term engagement. Hansen et al. [34] confirm that, by shifting towards diverse content, recommender systems may enhance users' satisfaction. Sguerra et al. [80] demonstrate that users' interests may evolve when exposed to new songs repeatedly. These off-line studies, together with our findings, suggest that while designing music recommender systems there should be a **tradeoff between diversification and repeated exposure** of unfamiliar music. Also considering the repetitive and sequential nature of music listening, the exploration of such tradeoff is a much-desired goal that the research community should aim at achieving.

Last, we decide to explore the impact of recommendations centering on a music genre that was unfamiliar and not known in-depth by the study participants. Indeed, to have listeners with a similar initial condition of *tabula rasa* minimizes the influence of preferences and prejudices, reducing the variation in terms of perceived diversity caused by prior experiences. We manipulated recommendation diversity to show different facets of a music genre to help the participants develop less-stereotypical visions of what was previously known probably only through mainstream artists. The function of recommendations in our study can be considered *educational*, where, through the calibration of diversity, we basically exposed participants to two different representations of the Electronic Music scene. Therefore, one of the main messages to take away from this study is that, through a thoughtful design, **recommender systems could be a vehicle for positive attitude change**. Under this lens, we connect with the idea discussed by Ferraro et al. [22] of designing recommender systems to enhance cultural citizenship, moving beyond the need for personalization often driven by commercial interests.

8 CONCLUSION

The impact assessment of music RS with regard to people's behavior, attitudes, habits, or beliefs is an active and challenging research topic that is attracting more and more practitioners from different disciplines. Until now, simulation methods have been the most-explored approach by the RS community to study the dynamics between users and items, and undoubtedly they have shed light on several behavioral aspects of such interaction. Nevertheless, the fact that *behaviorism is not enough* has already been pointed out by Ekstrand and Willemsen [19]. The focus only on what

users do, listen, or consume is indeed one of the main limitations we find in music RS research, which motivated us in designing this longitudinal study.

We did not target or expect to drastically change the listening habits of the study participants, because the function of RS is not, and should not be, to manipulate listeners towards specific artists or genres. Moreover, the socio-cultural background, empathetic and affective response, generational differences, music education, and many more aspects of life define what people listen to, and with this regard, people have many ways to discover and interact with music, and recommender systems are only one of those [10]. This also points to possible confounding factors that need to be carefully considered when designing a longitudinal user study. We intentionally focused on the impact of music recommendation on newcomers to a music genre, and we believe that the exploratory nature of this study has served the purpose of getting a better understanding of this phenomenon, while at the same time providing some directions for future explanatory research.

In conclusion, the extensive use of algorithmically driven recommendations, especially by young generations, raises questions in terms of human agency and autonomy of the next generation of music listeners [16], questions that recommender systems practitioners should not neglect.

9 REPRODUCIBILITY AND SUPPLEMENTARY MATERIALS

Supplementary materials are available in the online version of this article.

APPENDICES

A DATASET AND AUDIO MODELS

This appendix outlines the collection strategy of the material used to create the music recommendations for this study. Section A.1 describes the creation of the dataset, containing different kinds of EM. After selecting a pool of approximately 1,500 tracks from the dataset, several audio models were tested to understand what type of tracks' representation better served the purpose of diversifying the recommendations, as described in Section A.2.

A.1 Dataset

One of the goals of the study was to expose people not familiar with the EM culture to different genres that could fit under this label. Therefore, our aim was to select as many as possible varied tracks to represent the richness of this culture, even if we did not aim to create a dataset representing its entirety. To do that, we designed a semi-automatic method based on two sources: *Wikipedia* and *Every Noise at Once*.[14]

The initial step consisted in retrieving the list of EM genres from the Wikipedia page "List of electronic music genres" [97]. While several taxonomies and hierarchies of EM genres and subgenres have been published, we believe that the information on Wikipedia properly represents the variety of EM, enough precisely for our purposes. From that source, we obtained a set of **20 genres** and **320 subgenres**.[15] The second step was to map these subgenres to the ones part of the **Every Noise at Once (ENaO)** website. ENaO is described by its creator Glenn Mcdonald as "an ongoing attempt at an algorithmically-generated, readability-adjusted scatter-plot of the musical genre-space, based on data tracked and analysed for several genre-shaped distinctions by Spotify" [83]. Born as a debugging tool, in its current form, the website presents for each genre label a playlist formed by approximately 100 tracks, considered representative of the genre by

[14] https://everynoise.com

[15] We refer to *genre* when indicating the top-level genre labels in the Wikipedia hierarchy, highlighted in bold. Instead, we refer to the others as *subgenres*.

Spotify algorithms. Mapping Wikipedia to ENaO, we linked to the **20 genres** only **181 subgenres**, for which we retrieved the corresponding playlists.

A few aspects of the aforementioned approach should be noted: First, even if we are aware of the dynamical, intrinsic, ambiguous, and context-dependent nature of concepts such as genre and style [73], approaching music cultures as broad as Electronic Music is quite natural to identify different aesthetic and social characteristics in its several subgenres [61]. For instance, even without knowing anything about EM genres, it is possible to recognize the differences between the fast breakbeats of the *Drum and Bass* and the slow-tempo beats of *Downtempo*. As a matter of fact, in a previous study, we showed how familiarity with styles and subgenres highly influence the perceived diversity of people exposed to EM tracks [74]. Even if it could have left out some genres, pursuing a bottom-down approach to find representative Electronic Music, starting from genre labels arriving at tracks, seems to us the most obvious approach to creating a varied EM dataset.

Second, the choice of using ENaO playlists to select candidate tracks could be criticized because of the opaqueness of the Spotify algorithms. Indeed, it is not possible to find the exact description of the procedure that assigns a genre label to artists and albums. Even having in mind this limitation, after exploring in depth the ENaO website, we believe that it provides a good representation of the genres on its map, sharing the idea of ENaO creator that "the point of the map, as with the genres, is not to resolve disputes but to invite you to explore music" [60]. Under this lens, we do not argue that the classification in this study is the ultimate one, but one of the possible classifications that may help listeners navigate the EM culture.

After a preliminary exploration of the dataset, we performed a manual cleaning of some of the music selected to be sure that crossover genres would not enter into the final listening sessions. We do not argue if *Electronic Rock* may be considered *Electronic* music, *Rock* music, or both, but we believe that crossover music did not fit the purpose of our study. In the end, we restricted the dataset to **20 genres** (*ambient*, *bass music*, *breakbeat*, *chill out*, *disco*, *drum and bass*, *electroacoustic*, *electronica*, *garage*, *hardcore*, *hardstyle*, *hauntology*, *house*, *IDM*, *jungle*, *noise*, *plunderphonics*, *techno*, *trance*, *videogame*) and **165 subgenres** listed in Appedix B (Table 7). At that point, we had a pool of around 16k tracks (~100 tracks for subgenre) listenable on Spotify.

Further filtering was applied by looking at the popularity of the tracks. Indeed, several studies have proven the influence of familiarity on music preferences [13]. To avoid creating a *popularity effect* in our study, i.e., participants' ratings influenced by the popularity of a track, we filtered the dataset by using two indicators: (1) Spotify track' popularity; (2) YouTube view count. In detail, we implement the following procedure: From the 16k tracks for which we already had the Spotify ID, we filtered out the ones with popularity less than the first quartile Q1 and major than the third quartile Q3 computed on the Spotify popularity indicator. For the remaining ones, we made use of a Python wrapper to search with the YouTube API for the corresponding video by using the track names. As a result, we obtain a total of approximately 8k tracks with Spotify ID and YouTube ID, already filtered by Spotify popularity. The last step was to further filter out tracks according to the YouTube view count applying the same logic of Spotify popularity, including only tracks with views between Q1 and Q3. After the second filtering, we randomly chose 10 tracks for each subgenre obtaining **1,444 candidate tracks**, for which, finally, we extracted the audio embedding, as reported in the next section.

A.2 Audio Models

The use of deep representation models in MIR research is widespread and applied in several retrieval and classification tasks, e.g., auto-tagging, instrument recognition, genre classification, and ultimately music recommendations [100]. Nevertheless, the trustworthiness of such representations is still under the scrutiny of the research community, partly because of the still low

interpretability in comparison to traditional hand-crafted music representations, usually informed by human music domain knowledge [48]. Even if the goal of this study was not to perform a rigorous comparative analysis of different music representation models, we were nevertheless interested in creating diversity-aware content-based music recommendations, on the one hand, based on state-of-the-art deep learning models, and on the other hand, interpretable enough to make us aware of the characteristics underlying the diversification outcomes. Consequently, we first started exploring several deep representation models, and then we attempted to interpret those by using a set of hand-crafted features.

A.2.1 Deep Representation Models. We experimented with four audio representation models available in *Essentia* [9], an open-source library for audio and music analysis. Notably, Essentia provides Tensorflow deep learning models built on different architectures and trained with different datasets, open sources and publicly available [4]. Among those available, we tested the following models:

- **EffNet–Discogs:** EfficientNet [90] trained with the **Discogs-Effnet Dataset (DED)**.
- **MusicNN–MSD/MusicNN–MTT:** MusicNN [72] trained with the **Million Song Dataset (MSD)** [7] and the **MagnaTagATune (MTT)** dataset [52].
- **VGGish–AudioSet:** VGG [81] trained with the AudioSet dataset [29].

A detailed description of the models, datasets, and implementations can be found online,[16] but, here, we discuss a few aspects relevant to our study.

First, the compared models have been used as audio *end-to-end* learning systems, where, given the audio as input, they return the desired output (i.e., embedding) learned from data (i.e., the audio signal). We chose to not consider a multi-modal approach (e.g., including metadata) to focus mostly on the listenable differences of the tracks in the dataset. Second, even if we employed these models as feature extractors, i.e., to obtain an embedded representation of each audio in our dataset, their original intended purposes are quite different. For instance, *VGG* was proposed to tackle the task of image recognition, while *MusicNN* primarily focuses on auto-tagging. Therefore, our final choice was based not on the architecture that better performs according to its original scope, but on the feature extractor that better worked according to our objectives. Third, we intentionally selected a quite heterogeneous set of models, especially in terms of datasets used in the training stage. Indeed, *AudioSet* is formed by almost two million clips annotated using sound labels not always specific to music (e.g., footstep, bark, cutlery). Instead, *MSD* and *MTT* datasets are annotated with 50 tags describing the genre, instrumentation, or also mood of the tracks. Last, *DED* contains tracks annotated with 400 music styles according to the taxonomy of the crowdsourced database *Discogs*.

After selecting the four models, we used them to extract the audio embeddings for the 1,444 candidate tracks part of our dataset. In the former three cases, we obtained 200-dimensional embeddings, while from the *VGGish-AudioSet,* we got 1,024-dimensional embeddings. At that point, we were interested in a model that could coherently represent Electronic Music according to the genre labels that we already got. Indeed, assuming that the tracks representative of one genre, i.e., coming from the same *Every Noise at Once* playlist, should be more similar to one another in comparison to the tracks of another genre, our problem was translated into measuring how close the embeddings placed tracks from the same genre in the embedded space. Therefore, we used the tracks' genre labels to create 20 clusters, and then we measured the consistency of each cluster by performing a Silhouette analysis [78].

[16] https://essentia.upf.edu/models.html

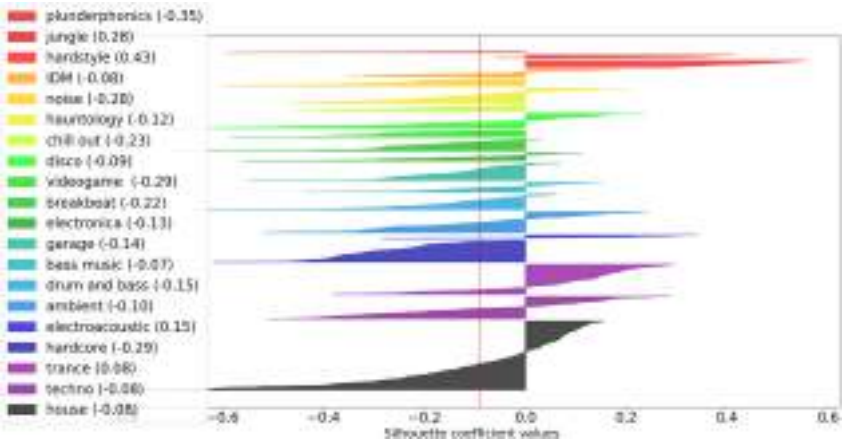

Fig. 12. Silhouette scores for the tracks clustered by genre. The dashed line indicates the average Silhouette score over all the genres. In the legend, the values in the parentheses indicate the average genre score.

In brief, Silhouette analysis measures how much an item of a cluster is similar to other items of its own cluster compared to the ones of other clusters. It ranges from −1 to 1, where negative values indicate that an item has been poorly clustered, while positive values mean that an item has been properly matched. Given the high dimensionality of the embeddings, we chose to use cosine similarity to measure the distance between items. An example of silhouette scores for the 20 genre clusters is reported in Figure 12, computed using the cosine distance between the *EffNet-Discogs* embeddings. The score of each genre is computed by averaging the scores of its corresponding tracks. Among the others, we see that *Jungle* tracks are well-clustered together (.28), while, on the contrary, *Hardcore* tracks seem not (−.29). Other genres such as *House* have almost half tracks well-clustered while the other half poorly, obtaining an average score around zero (−.08).

Silhouette analysis is commonly performed to validate the output of clustering algorithms, however, in our case, we employed it as a tool to validate which of the four models provides us with more consistent embeddings according to our "ground-truth" labels. Averaging the silhouette scores over all the 20 genres, we obtained the best score using the *EffNet-Discogs* model (−.09), followed by *VGGish-AudioSet* (−.11), and then *MusicNN* (*MSD*: −.16, *MTT*: −.17). While the difference between models seems not significant, we believe that the use of the Discogs dataset may have boosted the performance of the EfficientNet architecture in creating audio representations that better reflect differences between Electronic Music tracks. Indeed, born originally as an Electronic Music database, Discogs is a huge source of knowledge of such culture. To properly compare the different models, each architecture should have been trained with the Discogs database, a task left to practitioners interested in understanding better how deep representations behave with EM tracks. From now on, when mentioning embeddings, we implicitly refer to the ones generated by the **EffNet-Discogs** model.

Figure 13 displays a 2-dimensional representation of the embedded space obtained with this model. In the center of the circle, we notice a quite messy situation with tracks from different genres placed near each other. However, going near the border, we see a few clusters more defined, for instance, the *Trance* tracks on the right or *Drum and Bass* on the bottom. We see also how genres that shared music properties are clustered near each other, such as *Drum and Bass–Jungle*, or *Hardcore–Hardstyle*. From the scatter plot, we may also have an intuition about the logic behind the placing of the points in the space according to characteristics such as the BPM or the softness

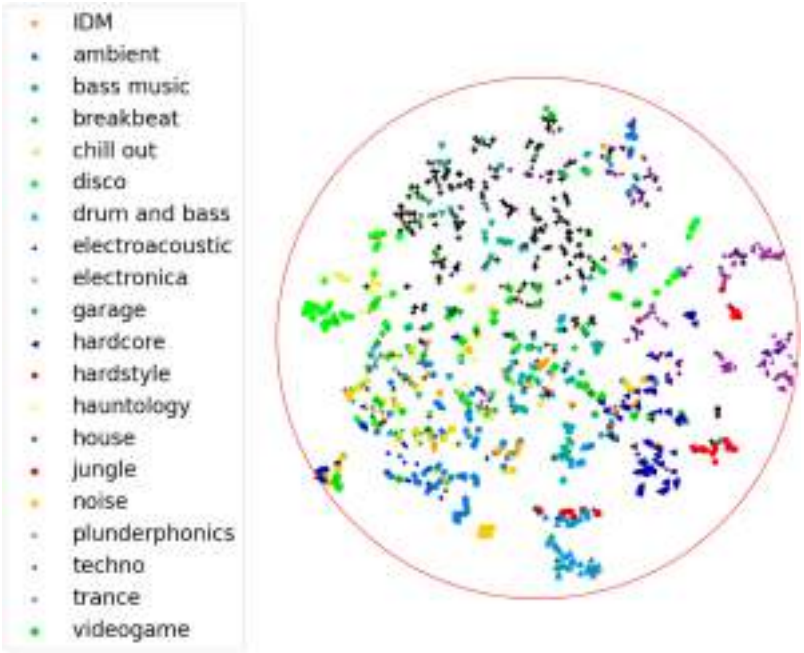

Fig. 13. Two-dimensional t-SNE projection of candidate tracks' embedding.

of the sounds. However, to interpret in depth the nature of the embeddings, in the next section, we continue our analysis focusing on a series of hand-crafted music features.

A.2.2 Hand-crafted Music Features. Following a first exploration of the embeddings, we scrutinized in depth their relationship with four music features: *tempo*, *danceability*, *acousticness,* and *instrumentalness*. The reason why we selected these features is twofold. First, the embeddings seemed to be in some way informative of the distribution of those. For instance, in the embedded space, the tempo apparently decreased going from bottom to top, while the danceability in the opposite way. Second, by using these features, we had the opportunity to verify the reliability of Essentia's feature extractor[17] and the Spotify API. Both present advantages and disadvantages, and according to the available resources, one could prefer one method rather than the other.

Indeed, a main drawback of Essentia is the need for the audio track file to extract the features, while using the Spotify API just with the track ID, it is possible to access several audio features. However, Spotify algorithms are proprietary, while Essentia being open source, it is possible to verify the exact functioning of the extraction process. Eventually, we checked the Pearson correlation coefficient (ρ) between the features extracted with Essentia and Spotify, and using the tracks in our dataset, we found a positive correlation: *tempo* ($\rho = .29, p < .01$), *instrumentalness* ($\rho = .42$, $p < .01$), *danceability* ($\rho = .50, p < .01$), and *acousticness* ($\rho = .61, p < .01$). Therefore, in terms of analysis, no particular difference should emerge if using one method rather than the other. From now on, when referring to features, we implicitly refer to the ones extracted with **Essentia**.

The track features' distribution (Figure 14) highlights some characteristics of Electronic Music in the dataset. First, most of the tracks have a tempo of between 120 and 150 BPM, with some outliers over 160 and under 90 BPM. It is worth noting that tempo estimation algorithms suffer

[17] https://essentia.upf.edu/streaming_extractor_music.html

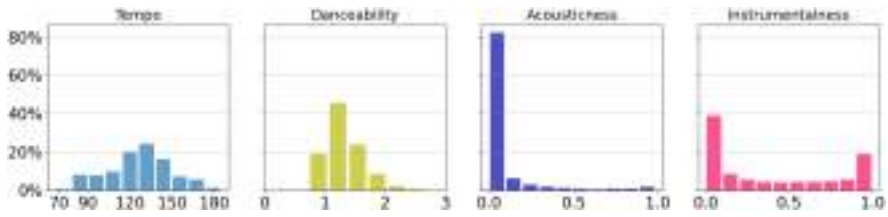

Fig. 14. Histograms of candidate tracks' feature distributions.

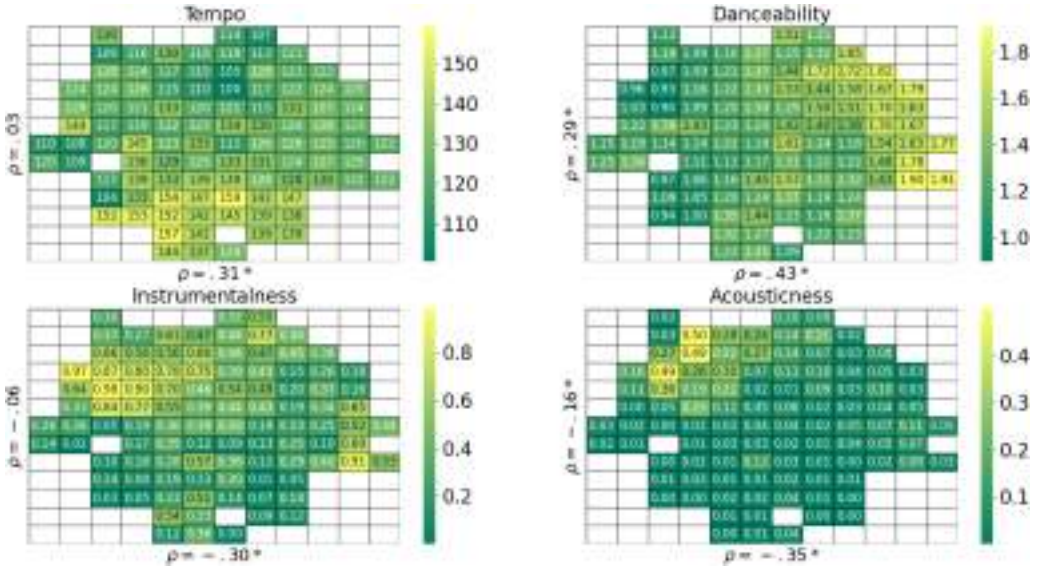

Fig. 15. Embeddings-features block distribution. The asterisk indicates correlation coefficients with $p < .001$. A positive correlation of the x-axis means that values increase while going from left to right, while in the y-axis from bottom to top.

from the problem of the so-called *octave errors,* i.e., assigning 80 instead of 160 BPM or vice versa [79], hence, some of these outliers could be a result of these errors. In terms of danceability, the extractor returns values between 0 and 3, where higher values mean more danceable. We have the majority of tracks with values between 1 and 2, with 20% scoring less than 1 and a small percentage scoring more than 2. Acousticness and instrumentalness are computed as probabilities ranging from 0 to 1. In the case of the former, 0 means almost certainly no presence of acoustic instruments, while 1 is the opposite scenario. For the latter, 0 means almost certainly the presence of a singing voice, while 1 the absence of singing voice parts. In our dataset, it is not surprising to observe that the majority of the tracks are classified as non-acoustic, while we see that the presence of singing voices is less skewed in comparison with the acousticness. Computing the correlation between the four features, we found that danceability is negatively correlated with both acousticness and instrumentalness ($\rho = -.44$, $p < .01$ and $\rho = -.66$, $p < .01$), meaning that the more danceable tracks in the dataset are the ones without acoustic instrumentation but with singing parts. Instead, these two latter features were positively correlated among them ($\rho = .31$, $p < .01$).

As the last step before defining the diversification strategy, we explored the link between the embeddings and the hand-crafted features. To do that, we first divided into equally sized blocks

the 2-dimensional projection of the embedded space created (see Figure 13). Then, we assigned each track to a block according to its position in the space. Finally, in each block, we averaged the feature values of its tracks. Because of the non-uniform density of the embedded space, tracks were not equally distributed among blocks. Figure 15 displays the four block-heatmaps linking the features to the embedding distribution.

We may observe how the embeddings coherently clustered the tracks with regard to the selected features. For instance, tracks with extreme tempos (more than 130 BPM and less than 110 BPM) are distributed at the bottom of the embedded space. On the contrary, tracks with high danceability are mostly in the left part of the heatmap. Acousticness and instrumentalness instead have higher values in the top-right corner. Further validation can be obtained by looking at the relationship between the heatmaps and the genre clusters. For instance, the *Drum and Bass*, *Hardcore*, and *Hardstyle* clusters located at the bottom of the plot correspond to the blocks where the tempo is higher. The *Techno*, *Trance,* and part of the *House* clusters on the left instead are located where the danceability is higher. Instrumentalness goes up in correspondence with *Disco* and *Hauntology* clusters, where also the acousticness is quite high.

B ADDITIONAL TABLES AND FIGURES

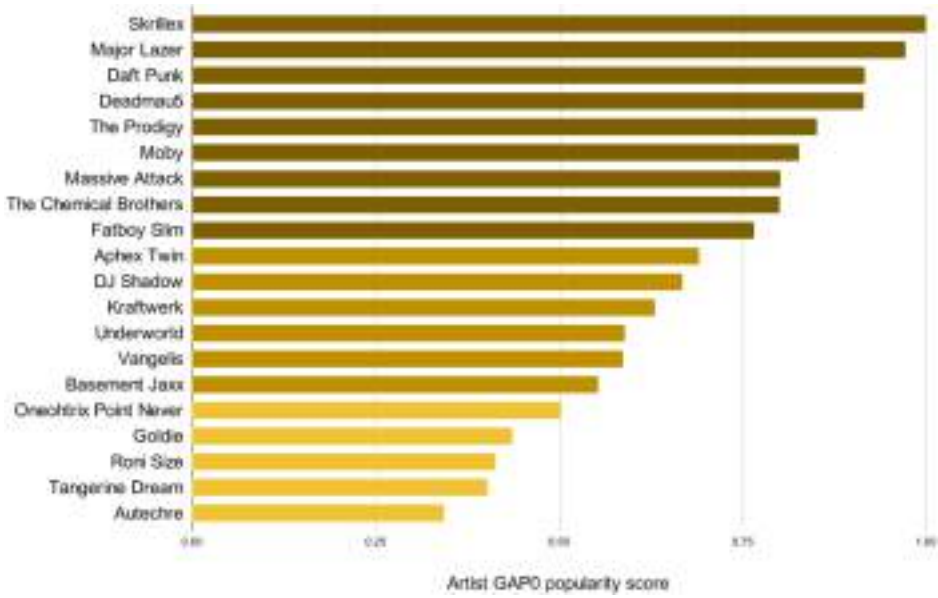

Fig. 16. List of artists and corresponding GAP0 score for assessing the participants' familiarity with EM.

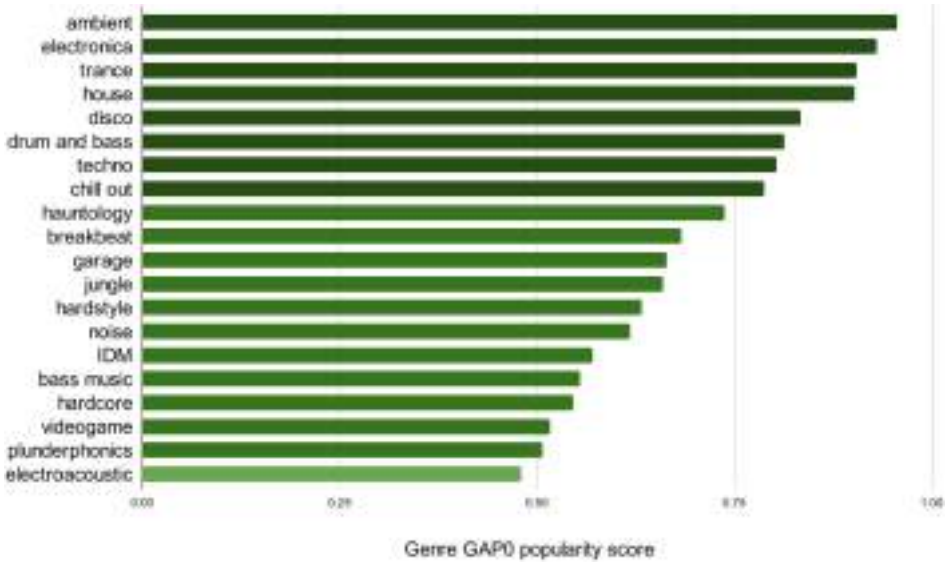

Fig. 17. List of genres and corresponding GAP0 score for assessing the participants' familiarity with EM.

Table 7. List of Genres and Subgenres Included in the Dataset

Genre	Subgenre(s)
ambient	ambient dub techno, ambient industrial, dark ambient, drone, dungeon synth, illbient, lowercase, new age, new isolationism, space ambient.
bass music	footwork, future bass, kawaii future bass, wave.
breakbeat	big beat, broken beat, hardcore breaks, jersey club, nu skool breaks, progressive breaks.
chill out	downtempo, psydub, trip hop.
disco	boogie, city pop, eurobeat, eurodance, hi-nrg, italo dance, nu disco, post-disco.
drum and bass	atmospheric dnb, darkstep, drumfunk, jump up, liquid funk, neurofunk, sambass.
electroacoustic	acousmatic, electroacoustic composition, electroacoustic improvisation, musique concrete.
electronica	berlin school, folktronica, jazztronica, livetronica.
garage	bassline, brostep, chillstep, dubstep, future garage, grime, speed garage, uk funky, uk garage, wonky.
hardcore	acidcore, breakcore, digital hardcore, doomcore, frenchcore, happy hardcore, industrial hardcore, j-core, makina, speedcore, terrorcore.
hardstyle	euphoric hardstyle, jumpstyle, rawstyle.
hauntology	chillwave, darksynth, future funk, hardvapour, sovietwave, synthwave, vaporwave.
house	acid house, afro house, amapiano, ambient house, bass house, brazilian bass, chicago house, complextro, deep euro house, deep house, disco house, diva house, dutch house, electro swing, fidget house, funky house, future house, garage house, ghettotech, gqom, hard bass, italo house, jazz house, kwaito, latin house, melbourne bounce, microhouse, moombahton, outsider house, progressive house, slap house, soulful house, tech house, tribal house.
IDM	drill and bass, glitch, glitch hop.
jungle	ragga jungle.
noise	death industrial, japanoise, power electronics, power noise.
plunderphonics	—
techno	acid techno, ambient techno, berlin minimal techno, bleep techno, detroit techno, dub techno, hard techno, industrial techno, minimal techno, raggatek, schranz.
trance	acid trance, dark psytrance, dream trance, full on, goa trance, hands up, hard trance, nitzhonot, progressive psytrance, progressive trance, psychedelic trance, suomisaundi, tech trance, uplifting trance, vocal trance.
videogame	bitpop, chiptune, nintendocore, skweee.

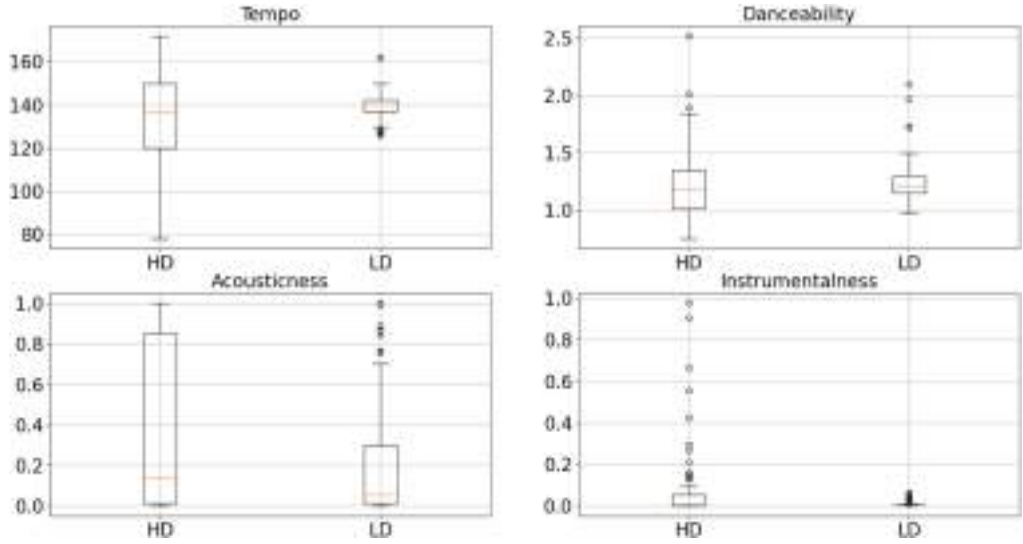

Fig. 18. Boxplots of the features' distribution for the high diversity (HD) and low diversity (LD) recommendation lists.

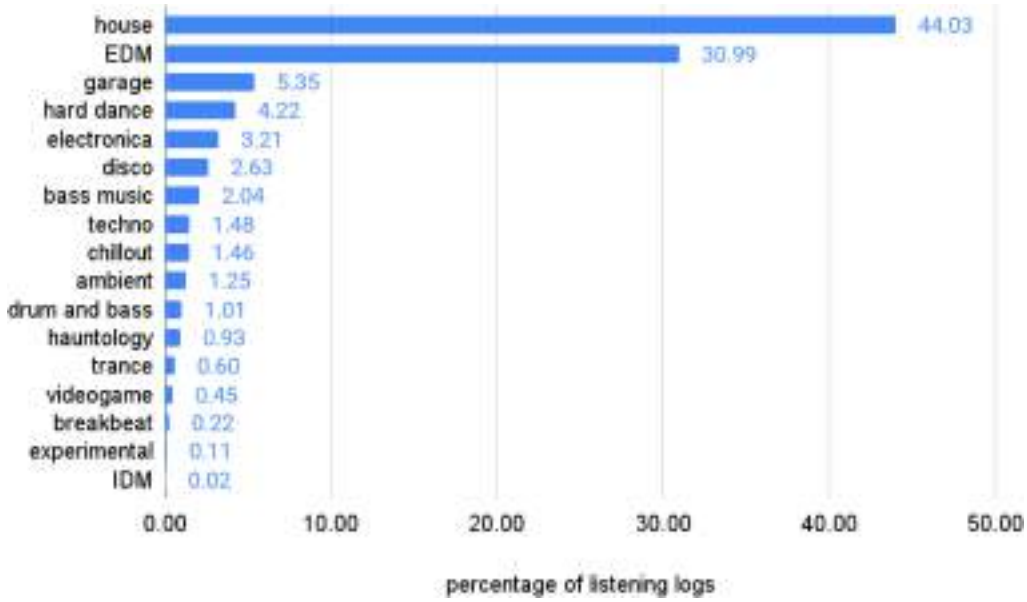

Fig. 19. Genres ranked by popularity in the study participants' listening logs.

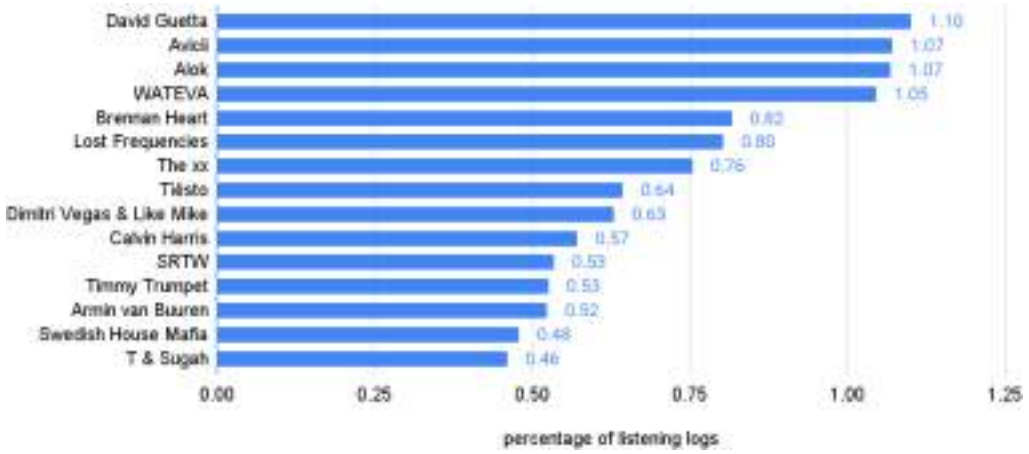

Fig. 20. Top artists ranked by popularity in the study participants' listening logs.

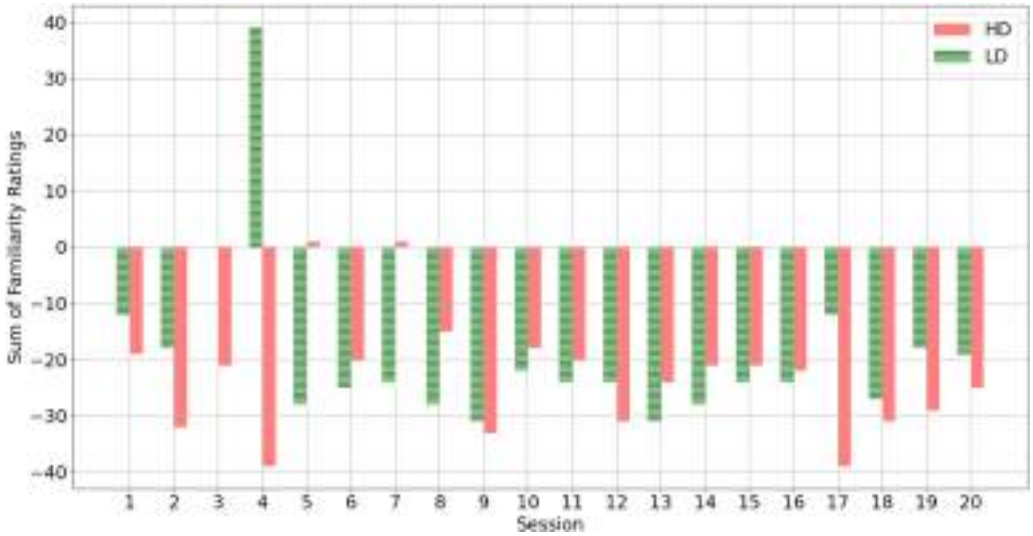

Fig. 21. Distribution of familiarity ratings.

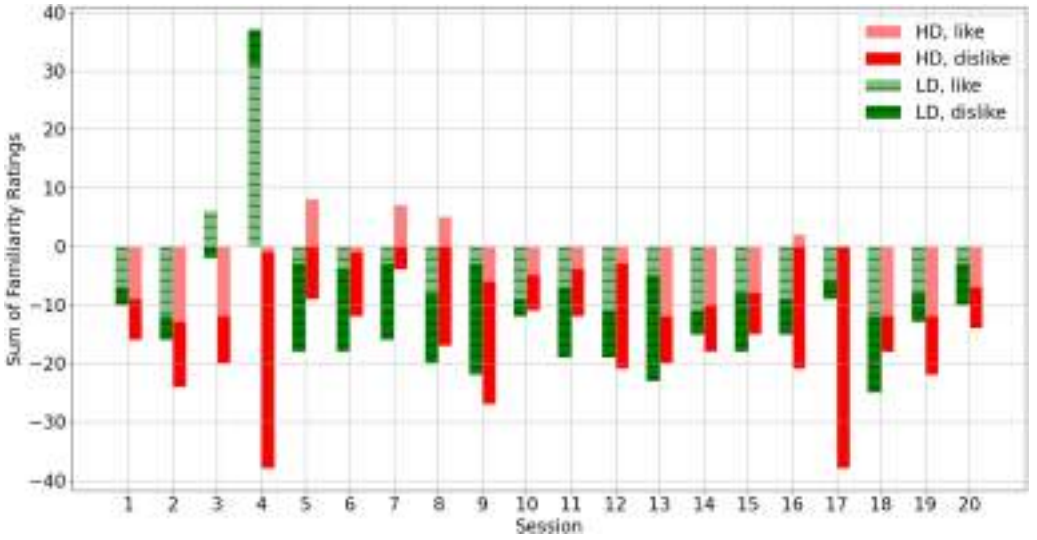

Fig. 22. Distribution of familiarity ratings split among participants who liked the session (*light bar*) and participants who disliked (*dark bar*).

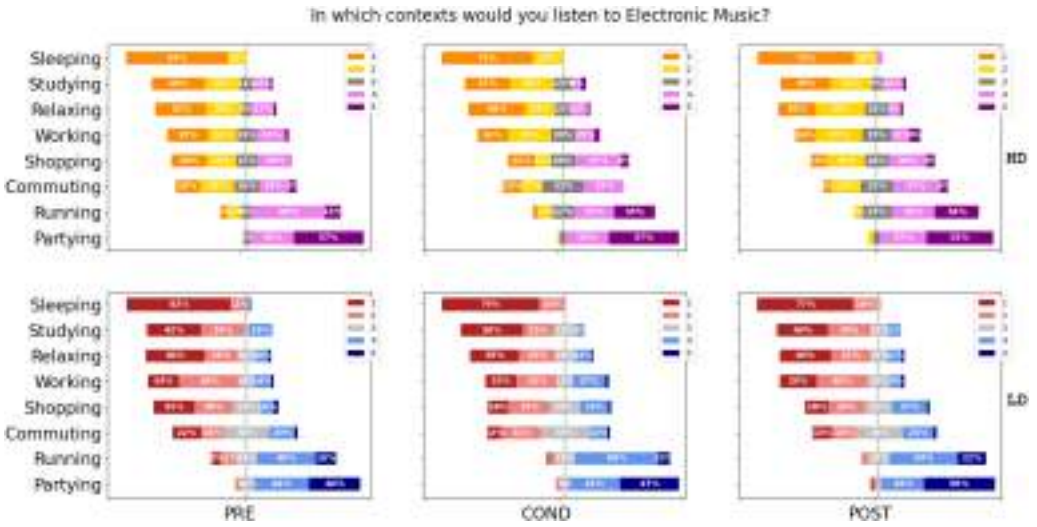

Fig. 23. Distribution of participants' ratings of the characteristics associated with listening contexts at the beginning of the experiment (*PRE*), after the exposure (*COND*), and at the end (*right*), separately for the HD group (*top*) and the LD group (*bottom*). In the legend, the values of the Likert-item selected are reported (1: Totally Disagree, 5: Totally agree).

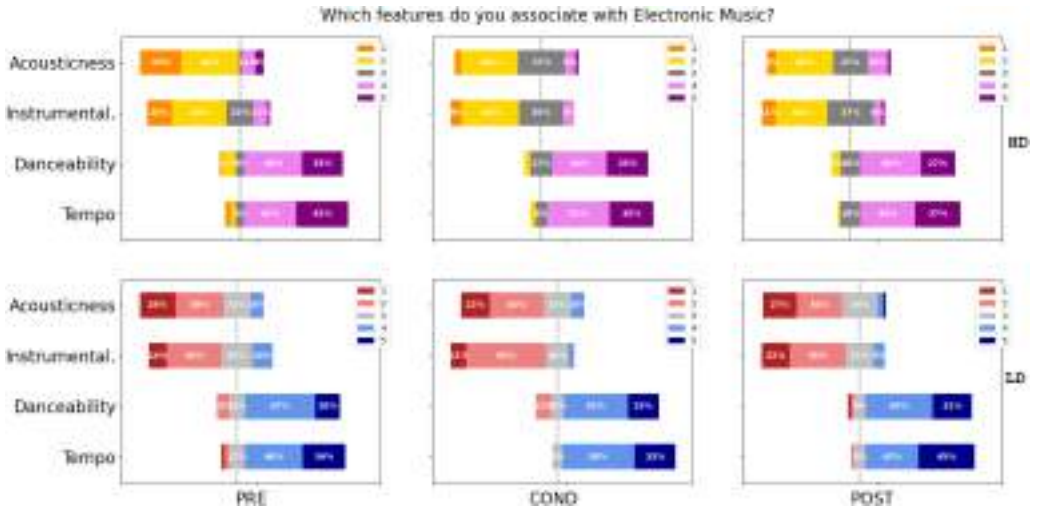

Fig. 24. Distribution of participants' ratings of the musical characteristics associated with EM tracks at the beginning of the experiment (*PRE*), after the exposure (*COND*), and at the end (*right*), separately for the HD group (*top*) and the LD group (*bottom)*. The values for the Likert items are: *acousticness* (1: mostly low, 5: mostly high), *instrumentalness* (1: mostly low, 5: mostly high), *danceability* (1: mostly low, 5: mostly high), *tempo* (1: mostly slow, 5: mostly fast).

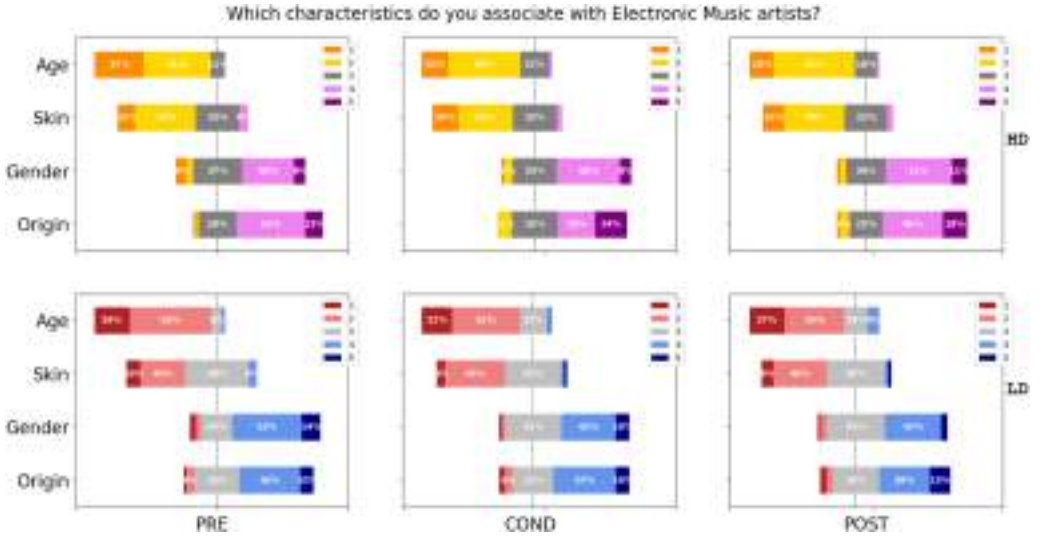

Fig. 25. Distribution of participants' ratings of the characteristics associated with EM artists at the beginning of the experiment (*PRE*), after the exposure (*COND*), and at the end (*right*), separately for the HD group (*top*) and the LD group (*bottom*). The values for the Likert items are: *age* (0: mostly under 40, 5: mostly over 40), *skin* (0: mostly white-skinned, 5: mostly dark-skinned), *gender* (0: mostly women or other gender minorities, 5: mostly men); *origin* (0: mostly low income/developing countries, 5: mostly high income/developed countries).

REFERENCES

[1] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Inf. Syst. Res.* 24, 4 (2013), 956–975. DOI: https://doi.org/10.1287/isre.2013.0497

[2] Gediminas Adomavicius, Dietmar Jannach, Stephan Leitner, and Jingjing Zhang. 2021. Understanding longitudinal dynamics of recommender systems with agent-based modeling and simulation. In *Workshop on Simulation Methods for Recommender Systems, co-located with ACM RecSys*. 21–24.

[3] AllMusic. 2022. Electronic Artists Highlights. Retrieved from https://www.allmusic.com/genre/electronic-ma0000002572/artists

[4] Pablo Alonso-Jiménez, Dmitry Bogdanov, Jordi Pons, and Xavier Serra. 2020. Tensorflowaudio models in Essentia. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20)*. 266–270. DOI: https://doi.org/10.1109/ICASSP40776.2020.9054688

[5] Ashton Anderson, Lucas Maystre, Rishabh Mehrotra, Ian Anderson, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on Spotify. In *Web Conference*. 2155–2165. DOI: https://doi.org/10.1145/3366423.3380281

[6] Guy Aridor, Duarte Goncalves, and Shan Sikdar. 2020. Deconstructing the filter bubble: User decision-making and recommender systems. In *14th ACM Conference on Recommender Systems (RecSys'20)*. Association for Computing Machinery, New York, NY, 82–91. DOI: https://doi.org/10.1145/3383313.3412246

[7] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *12th International Society for Music Information Retrieval Conference*. 591–596. DOI: https://doi.org/10.7916/D8NZ8J07

[8] Rachel M. Bittner, Eric J. Humphrey, and Juan Pablo Bello. 2016. pysox: Leveraging the audio signal processing power of SoX in Python. In *17th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*.

[9] 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. *14th International Society for Music Information Retrieval Conference (ISMIR'13)*. 493–498.

[10] Georgina Born, Jeremy Morris, Fernando Diaz, and Ashton Anderson. 2021. *Artificial Intelligence, Music Recommendation, and the Curation of Culture*. Technical Report. Schwartz Reisman Institute for Technology and Society.

[11] Kelly Caine. 2016. Local standards for sample size at CHI. In *CHI Conference on Human Factors in Computing Systems (CHI'16)*. Association for Computing Machinery, New York, NY, 981–992. DOI: https://doi.org/10.1145/2858036.2858498

[12] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *12th ACM Conference on Recommender Systems (RecSys'18)*. Association for Computing Machinery, New York, NY, 224–232. DOI: https://doi.org/10.1145/3240323.3240370

[13] Anthony Chmiel and Emery Schubert. 2017. Back to the inverted-U for music preference: A review of the literature. *Psychol. Music* 45, 6 (2017), 886–909. DOI: https://doi.org/10.1177/0305735617697507

[14] Eric Clarke, Tia DeNora, and Jonna Vuoskoski. 2015. Music, empathy and cultural understanding. *Physics. Life Rev.* 15 (2015), 61–88. DOI: https://doi.org/10.1016/j.plrev.2015.09.001

[15] Jacob Cohen. 1992. A power primer. *Quantit. Meth. Psychol.* 112, 1 (1992), 155–159.

[16] David Crider. 2022. Listening, but not being heard: Young women, popular music, streaming, and radio. *Pop. Music Soc.* 45, 5 (2022), 1–17. DOI: https://doi.org/10.1080/03007766.2022.2111513

[17] Edith D. de Leeuw and Peter Lugtig. 2015. Dropouts in longitudinal surveys. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 1–6. DOI: https://doi.org/10.1002/9781118445112.stat06661.pub2

[18] Michael D. Ekstrand, Allison Chaney, Pablo Castells, Robin Burke, David Rohde, and Manel Slokom. 2021. SimuRec: Workshop on synthetic data and simulation methods for recommender systems research. In *15th ACM Conference on Recommender Systems (RecSys'21)*. Association for Computing Machinery, New York, NY, 803–805. DOI: https://doi.org/10.1145/3460231.3470938

[19] Michael D. Ekstrand and Martijn C. Willemsen. 2016. Behaviorism is not enough: Better recommendations through listening to users. In *10th ACM Conference on Recommender Systems (RecSys'16)*. Association for Computing Machinery, New York, NY, 221–224. DOI: https://doi.org/10.1145/2959100.2959179

[20] Francesco Fabbri, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. 2022. Rewiring what-to-watch-next recommendations to reduce radicalization pathways. In *ACM Web Conference (WWW'22)*. Association for Computing Machinery, New York, NY, 2719–2728. DOI: https://doi.org/10.1145/3485447.3512143

[21] Andrés Ferraro. 2021. *Music Recommender Systems: Taking into Account the Artists' Perspective*. Ph. D. Dissertation. Universitat Pompeu Fabra.

[22] Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2022. Measuring commonality in recommendation of cultural content: Recommender systems to enhance cultural citizenship. In *16th ACM Conference*

on Recommender Systems (RecSys'22). Association for Computing Machinery, New York, NY, 567–572. DOI: https://doi.org/10.1145/3523227.3551476

[23] Andrés Ferraro, Dietmar Jannach, and Xavier Serra. 2020. Exploring longitudinal effects of session-based recommendations. In *14th ACM Conference on Recommender Systems*. 474–479. DOI: https://doi.org/10.1145/3383313.3412213

[24] Daniel Fleder and Kartik Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Manag. Sci.* 55, 5 (2009), 697–712. DOI: https://doi.org/10.1287/mnsc.1080.0974

[25] Daniel M. Fleder and Kartik Hosanagar. 2007. Recommender systems and their impact on sales diversity. In *8th Annual Conference on Electronic Commerce*. 192–199. DOI: https://doi.org/10.1145/1250910.1250939

[26] Carina Freitas, Enrica Manzato, Alessandra Burini, Margot J. Taylor, Jason P. Lerch, and Evdokia Anagnostou. 2018. Neural correlates of familiarity in music listening: A systematic review and a neuroimaging meta-analysis. *Front. Neurosci.* 12, Oct. (2018), 1–14. DOI: https://doi.org/10.3389/fnins.2018.00686

[27] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and evaluating user satisfaction with music discovery. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*. Association for Computing Machinery, New York, NY, 55–64. DOI: https://doi.org/10.1145/3209978.3210049

[28] Bertram Gawronski, Mike Morrison, Curtis E. Phills, and Silvia Galdi. 2017. Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personal. Soc. Psychol. Bull.* 43, 3 (2017), 300–312. DOI: https://doi.org/10.1177/0146167216684131

[29] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. 776–780.

[30] Anthony G. Greenwald, Debbie E. Mcghee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *J. Personal. Soc. Psychol.* 74, 6 (1998), 1464–1480.

[31] Tobias Greitemeyer, Jack Hollingdale, and Eva Traut-Mattausch. 2015. Changing the track in music and misogyny: Listening to music with pro-equality lyrics improves attitudes and behavior toward women. *Psychol. Pop. Media Cult.* 4, 1 (2015), 56–67. DOI: https://doi.org/10.1037/a0030689

[32] Tobias Greitemeyer and Anne Schwab. 2014. Employing music exposure to reduce prejudice and discrimination. *Aggress. Behav.* 40, 6 (2014), 542–551. DOI: https://doi.org/10.1002/ab.21531

[33] Louis Guttman. 1944. A basis for scaling qualitative data. *Amer. Sociol. Rev.* 9, 2 (1944), 139–150.

[34] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting consumption towards diverse content on music streaming platforms. In *14th ACM International Conference on Web Search and Data Mining (WSDM'21)*. Association for Computing Machinery, New York, NY, 238–246. DOI: https://doi.org/10.1145/3437963.3441775

[35] Md Rajibul Hasan, Ashish Kumar Jha, and Yi Liu. 2018. Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives. *Comput. Hum. Behav.* 80 (2018), 220–228. DOI: https://doi.org/10.1016/j.chb.2017.11.020

[36] Hanna Hauptmann, Nadja Leipold, Mira Madenach, Monika Wintergerst, Martin Lurz, Georg Groh, Markus Böhm, Kurt Gedrich, and Helmut Krcmar. 2021. Effects and challenges of using a nutrition assistance system: Results of a long-term mixed-method study. *User Model. User-adapt. Interact.* 32, 5 (2021). DOI: https://doi.org/10.1007/s11257-021-09301-y

[37] Naieme Hazrati, Mehdi Elahi, and Francesco Ricci. 2020. Simulating the impact of recommender systems on the evolution of collective users' choices. In *31st ACM Conference on Hypertext and Social Media (HT'20)*. 207–212. DOI: https://doi.org/10.1145/3372923.3404812

[38] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Inf., Commun. Soc.* 21, 2 (2018), 191–207. DOI: https://doi.org/10.1080/1369118X.2016.1271900

[39] IAIA, International Association for Impact Assessment. 2009. What Is IA. Retrieved from https://www.iaia.org/uploads/pdf/What_is_IA_web.pdf

[40] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara fallacy: Toward more impactful recommender systems research. *AI Mag.* 41, 4 (2020), 79–95. DOI: https://doi.org/10.1609/aimag.v41i4.5312

[41] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. *User Model. User-adapt. Interact.* 25, 5 (2015), 427–491. DOI: https://doi.org/10.1007/s11257-015-9165-3

[42] Dietmar Jannach, Oren Sar Shalom, and Joseph A. Konstan. 2019. Towards more impactful recommender systems research. In *ImpactRS Workshop, co-located with the 13th ACM Conference on Recommender Systems (RecSys'19)*. 15–17.

[43] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *AAAI/ACM Conference on AI, Ethics, and Society*. 383–390. DOI: https://doi.org/10.1145/3306618.3314288

[44] Rebecca R. Johnston and Gina M. Childers. 2022. Musical pantophagy: Is there an effect on changes in preference resulting from repeated exposure? *Psychol. Music* 50, 1 (2022), 141–158. DOI: https://doi.org/10.1177/0305735620987299
[45] Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. Association for Computing Machinery, New York, NY, 1105–1114. DOI: https://doi.org/10.1145/2207676.2208557
[46] Andrew Karpinski and Ross B. Steinman. 2006. The single category implicit association test as a measure of implicit social cognition. *J. Personal. Soc. Psychol.* 91, 1 (2006), 16–32. DOI: https://doi.org/10.1037/0022-3514.91.1.16
[47] Deanna Kemp and Frank Vanclay. 2013. Human rights and impact assessment: Clarifying the connections in practice. *Impact Assess. Proj. Apprais.* 31, 2 (2013), 86–96. DOI: https://doi.org/10.1080/14615517.2013.782978
[48] Jaehun Kim, Julián Urbano, Cynthia C. S. Liem, and Alan Hanjalic. 2019. Are nearby neighbors relatives? Testing deep music embeddings. *Front. Appl. Math. Statist.* 5, Nov. (2019), 1–17. DOI: https://doi.org/10.3389/fams.2019.00053
[49] Emily Jane Kothe and Mathew Ling. 2019. Retention of participants recruited to a multi-year longitudinal study via Prolific. *PsyArXiv Preprints* (2019), 1–8. https://doi.org/10.31234/osf.io/5yv2u
[50] Christos Koutlis, Manos Schinas, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2020. GAP: Geometric aggregation of popularity metrics. *Information* 11, 6 (2020). DOI: https://doi.org/10.3390/INFO11060323
[51] Reinhard Kreissl, Florian Fritz, and Lars Ostermeier. 2015. Societal impact assessment. In *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed.), James D. Wright (Ed.). Elsevier, Oxford, 873–877. DOI: https://doi.org/10.1016/B978-0-08-097086-8.10561-6
[52] Edith Law and Luis von Ahn. 2009. Input-agreement: A new mechanism for collecting data using human computation games. In *SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. Association for Computing Machinery, New York, NY, 1197–1206. DOI: https://doi.org/10.1145/1518701.1518881
[53] Dokyun Lee and Kartik Hosanagar. 2014. Impact of recommender systems on sales volume and diversity. In *35th International Conference on Information Systems (ICIS'14)*. 1–15.
[54] Dokyun Lee and Kartik Hosanagar. 2017. How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. *Inf. Syst. Res.* 30, 1 (2017), 239–259. DOI: https://doi.org/10.1287/isre.2018.0800
[55] Yu Liang and Martijn C. Willemsen. 2022. Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences. In *16th ACM Conference on Recommender Systems (RecSys'22)*. Association for Computing Machinery, New York, NY, 3–13. DOI: https://doi.org/10.1145/3523227.3546772
[56] Yu Liang and Martijn C. Willemsen. 2022. Promoting music exploration through personalized nudging in a genre exploration recommender. *Int. J. Hum.-comput. Interact.* 39, 7 (2022), 1–24. DOI: https://doi.org/10.1080/10447318.2022.2108060
[57] Luminate. 2022. Spotlight On: Electronic/Dance Music. Retrieved from https://luminatedata.com/reports/luminate-releases-new-genre-insights
[58] Guy Madison and Gunilla Schiölde. 2017. Repeated listening increases the liking for music regardless of its complexity: Implications for the appreciation and aesthetics of music. *Front. Neurosci.* 11, Mar. (2017), 1–13. DOI: https://doi.org/10.3389/fnins.2017.00147
[59] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *29th ACM International Conference on Information & Knowledge Management (CIKM'20)*. Association for Computing Machinery, New York, NY, 2145–2148. DOI: https://doi.org/10.1145/3340531.3412152
[60] Glenn McDonald. 2013. How We Understand Music Genres. Retrieved from https://everynoise.com/EverynoiseIntro.pdf
[61] Kembrew McLeod. 2001. Genres, subgenres, sub-subgenres and more: Musical and social differentiation within electronic/dance music communities. *J. Pop. Music Stud.* 13, 1 (2001), 59–75. DOI: https://doi.org/10.1080/15242220131707165 1
[62] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. Association for Computing Machinery, New York, NY, 735–746. DOI: https://doi.org/10.1145/3442188.3445935
[63] Richard K. Morgan. 2012. Environmental impact assessment: The state of the art. *Impact Assess. Proj. Apprais.* 30, 1 (2012), 5–14. DOI: https://doi.org/10.1080/14615517.2012.661557
[64] Fabio Morreale. 2021. Where does the buck stop? Ethical and political issues with AI in music creation. *Trans. Int. Soc. Music Inf. Retriev.* 4 (2021), 105–113.
[65] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Data & Society. Retrieved from https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf

[66] Felipe L. Navarro, Perfecto Herrera, and Emilia Gómez. 2014. There are places I remember: personalized automatic creation of music playlists for Alzheimer's patients. *Poster presented at: The Neurosciences and Music V: cognitive stimulation and rehabilitation; 2014 May 29 - Jun 1*, Dijon. http://hdl.handle.net/10230/44908

[67] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *23rd International Conference on World Wide Web*. 677–686. DOI: https://doi.org/10.1145/2566486.2568012

[68] Nielsen. 2014. Who is the Electronic Music Listener? Retrieved from https://www.nielsen.com/insights/2014/who-is-the-electronic-music-listener

[69] Raphaël Nowak. 2016. When is a discovery? The affective dimensions of discovery in music consumption. *Pop. Commun.* 14, 3 (2016), 137–145. DOI: https://doi.org/10.1080/15405702.2016.1193182

[70] Adriana Partal and Kim Dunphy. 2016. Cultural impact assessment: A systematic literature review of current methods and practice around the world. *Impact Assess. Proj. Apprais.* 34, 1 (2016), 1–13. DOI: https://doi.org/10.1080/14615517.2015.1077600

[71] Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: A critical review, challenges, and future directions. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT'22)*. Association for Computing Machinery, New York, NY, 1929–1942. DOI: https://doi.org/10.1145/3531146.3533238

[72] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. 2018. End-to-end learning for music audio tagging at scale. In *19th International Society for Music Information Retrieval Conference (ISMIR'18)*.

[73] Lorenzo Porcaro, Carlos Castillo, and Emilia Gómez. 2021. Diversity by design in music recommender systems. *Trans. Int. Soc. Music Inf. Retr.* 4, 1 (2021), 114–126. DOI: https://doi.org/10.5334/tismir.106

[74] Lorenzo Porcaro, Emilia Gomez, and Carlos Castillo. 2022. Diversity in the music listening experience: Insights from focus group interviews. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'22)*. Association for Computing Machinery, New York, NY, 272–276. DOI: https://doi.org/10.1145/3498366.3505778

[75] Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2022. Perceptions of diversity in electronic music: The impact of listener, artist, and track characteristics. *Proc. ACM Hum-Comput. Interact.* 6, CSCW1, Article 109 (Apr. 2022), 26 pages. DOI: https://doi.org/10.1145/3512956

[76] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute. Retrieved from https://ainowinstitute.org/aiareport2018.pdf

[77] Peter J. Rentfrow and Samuel D. Gosling. 2007. The content and validity of music-genre stereotypes among college students. *Psychol. Music* 35, 2 (2007), 306–326. DOI: https://doi.org/10.1177/0305735607070382

[78] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7

[79] Hendrik Schreiber, Julián Urbano, and Meinard Müller. 2020. Music tempo estimation: Are we done yet? *Trans. Int. Soc. Music Inf. Retr.* 3, 1 (2020), 111. DOI: https://doi.org/10.5334/tismir.43

[80] Bruno Sguerra, Viet-Anh Tran, and Romain Hennequin. 2022. Discovery dynamics: Leveraging repeated exposure for user and music characterization. In *16th ACM Conference on Recommender Systems (RecSys'22)*. Association for Computing Machinery, New York, NY, 556–561. DOI: https://doi.org/10.1145/3523227.3551474

[81] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*.

[82] Maria Do Rosário Sousa, Félix Neto, and Etienne Mullet. 2005. Can music change ethnic attitudes among children? *Psychol. Music* 3, 33 (2005), 304–316. DOI: https://doi.org/10.1177/0305735605053735

[83] Spotify. 2018. How Spotify Discovers the Genres of Tomorrow. Retrieved from https://artists.spotify.com/blog/how-spotify-discovers-the-genres-of-tomorrow

[84] Spotify. 2022. Lessons Learned from Algorithmic Impact Assessments in Practice. Retrieved from https://engineering.atspotify.com/2022/09/lessons-learned-from-algorithmic-impact-assessments-in-practice/

[85] Alain Starke, Martijn Willemsen, and Chris Snijders. 2021. Promoting energy-efficient behavior by depicting social norms in a recommender interface. *ACM Trans. Interact. Intell. Syst.* 11, 3-4 (2021), 1–32. DOI: https://doi.org/10.1145/3460005

[86] Gijsbert Stoet. 2010. PsyToolkit: A software package for programming psychological experiments using Linux. *Behav. Res. Meth.* 42, 4 (2010), 1096–1104. DOI: https://doi.org/10.3758/BRM.42.4.1096

[87] Gijsbert Stoet. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teach. Psychol.* 44, 1 (2017), 24–31. DOI: https://doi.org/10.1177/0098628316677643

[88] K. P. Suresh. 2011. An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *J. Hum. Reproduct. Sci.* 4, 1 (2011), 8–11. DOI: https://doi.org/10.4103/0974-1208.82352

[89] Karl K. Szpunar, E. Glenn Schellenberg, and Patricia Pliner. 2004. Liking and memory for musical stimuli as a function of exposure. *J. Experim. Psychol.: Learn. Mem. Cogn.* 30, 2 (2004), 370–381. DOI: https://doi.org/10.1037/0278-7393.30.2.370

[90] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th International Conference on Machine Learning (Proceedings of Machine Learning Research,* Vol. 97), Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 6105–6114. Retrieved from https://proceedings.mlr.press/v97/tan19a.html

[91] Ming Tu. 2009. *The Effects of a Chinese Music Curriculum on Cultural Attitudes, Tonal Discrimination, Singing Accuracy, and Acquisition of Chinese Lyrics for Third-, Fourth-, and Fifth-grade Students.* Ph. D. Dissertation. University of Miami. Retrieved from http://search.proquest.com/docview/304932027?accountid=13155

[92] Frank Vanclay. 2003. International principles for social impact assessment. *Impact Assess. Proj. Apprais.* 21, 1 (2003), 5–12. DOI: https://doi.org/10.3152/147154603781766491

[93] Frank Vanclay and Ana Maria Esteves. 2011. Chapter 1 current issues and trends in social impact assessment. In *New Directions in Social Impact Assessment.* Edward Elgar Publishing, Cheltenham. Retrieved Jul 20, 2023, from https://doi.org/10.4337/9781781001196.00012

[94] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO'21).* Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3465416.3483294

[95] Oscar Voigt. 2016. EDM Festival Market Segmentation. Retrieved from https://www.linkedin.com/pulse/edm-festival-market-segmentation-euphoria-oscar-voigt/

[96] Jonna K. Vuoskoski, Eric F. Clarke, and Tia Denora. 2017. Music listening evokes implicit affiliation. *Psychol. Music* 45, 4 (2017), 584–599. DOI: https://doi.org/10.1177/0305735616680289

[97] Wikipedia. 2022. List of electronic music genres. Retrieved from https://en.wikipedia.org/wiki/List_of_electronic_music_genres

[98] Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. 2016. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Model. User-adapt. Interact.* 26, 4 (2016), 347–389. DOI: https://doi.org/10.1007/s11257-016-9178-6

[99] Amy A. Winecoff, Matthew Sun, Eli Lucherini, and Arvind Narayanan. 2021. Simulation as experiment: An empirical critique of simulation research on recommender systems. *ArXiv* (2021). arXiv:2107.14333

[100] Minz Won. 2022. *Representation Learning for Music Classification and Retrieval: Bridging the Gap between Natural Language and Music Semantics.* Ph. D. Dissertation. Universitat Pompeu Fabra.

[101] YouTube. 2022. How engagement metrics are counted. Retrieved from https://support.google.com/youtube/answer/2991785

[102] Robert B. Zajonc. 1980. Feeling and thinking: Preferences need no inferences. *Amer. Psychol.* 35, 2 (1980), 151.

[103] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Inf. Syst. Res.* 31, 1 (2020), 76–101. DOI: https://doi.org/10.1287/ISRE.2019.0876

[104] Meizi Zhou, Jingjng Zhang, and Gediminas Adomavicius. 2023. *Longitudinal Impact of Preference Biases on Recommender Systems' Performance.* Kelley School of Business Research Paper No. 2021-10. http://dx.doi.org/10.2139/ssrn.3799525

Received 30 November 2022; revised 13 May 2023; accepted 3 July 2023