



A Meta-Methodology for User Evaluation of Artificial Intelligence Generated Music; Using the Analytical Hierarchy Process, Likert and Emotional State Estimations

Miguel Civit, Véronique Draï-Zerbib, Francisco Cuadrado & Maria J. Escalona

To cite this article: Miguel Civit, Véronique Draï-Zerbib, Francisco Cuadrado & Maria J. Escalona (24 Mar 2025): A Meta-Methodology for User Evaluation of Artificial Intelligence Generated Music; Using the Analytical Hierarchy Process, Likert and Emotional State Estimations, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2025.2478265](https://doi.org/10.1080/10447318.2025.2478265)

To link to this article: <https://doi.org/10.1080/10447318.2025.2478265>



Published online: 24 Mar 2025.



Submit your article to this journal [↗](#)



Article views: 31



View related articles [↗](#)



View Crossmark data [↗](#)



A Meta-Methodology for User Evaluation of Artificial Intelligence Generated Music; Using the Analytical Hierarchy Process, Likert and Emotional State Estimations

Miguel Civit^a , Véronique Drai-Zerbib^b , Francisco Cuadrado^a , and Maria J. Escalona^c 

^aDepartment of Communication and Education, Universidad Loyola Andalucía, Seville, Spain; ^bLEAD – CNRS UMR5022 Université Bourgogne Institut Marey, Dijon, France; ^cDepartment of Computer Languages and Systems, E.T.S. Ingeniería Informática, Avda. Reina Mercedes s/n, Universidad de Sevilla, Seville, Spain

ABSTRACT

Artificial Intelligence (AI) music generation is a trending field, and many different generators are currently under development. However, no standardized evaluation method exists that can help researchers evaluate and compare AI-based music tools. To create a meta-methodology for AI music assessment based on user evaluation, that can be both standardized and deployed as a tailored implementation model adapted to the idiosyncrasies of specific generators and their intended applications, thereby helping future researchers draw comparisons between different systems. Two different decision trees/matrices are proposed to help researchers tailor their specific evaluation studies. As evaluation tools, the paper explores Likert and analytical hierarchy process (AHP) based surveys and emotional state estimations using facial action units, self-assessment, and physiological signals. A proof-of-concept study demonstrates the viability of the proposed tools for user-based AI music generation evaluation studies. A preference for audio music over symbolic music generation was observed, and this will require future research. The implementation of the proposed methodology and tools across the field will be helpful when comparing different systems in future research and to save time in the development of user-based studies. User-based evaluation studies are needed to prevent biases from passing into future iterations of AI music generators.

KEYWORDS

Automatic music generation; assisted music composition; artificial intelligence; user evaluation; evaluation studies

1. Introduction

Automatic music generation (AMG) is the use of artificial intelligence and machine learning techniques to create music with little to no human intervention. Many techniques have been used for AMG (Civit et al., 2022; García-Peñalvo & Vázquez-Ingelmo, 2023). Some of the most common are: probabilistic methods, such as Markov models, that generate music based on the probability distribution of musical elements in a given dataset; rule-based approaches, that use predefined algorithms or grammars to generate music and which often incorporate music theory principles; and deep learning models. Thanks to advances in this technology, it has been possible to develop different music generation models, including recurrent neural networks (RNNs), long-short-term memory (LSTM) networks, generative adversarial networks (GANs) and, in recent years, transformer-based or large language model (LLM) architectures.

These generation systems can create music either in the symbolic domain, i.e., the system output is equivalent to a musical score or in raw audio format, where the output is produced directly as a sound file. Until very recently, most generators were symbolic, and they produced outputs of higher quality than those created by audio-based generators.

With the use of high-performance encoders and better language models, audio-based music generators have become a major trending alternative.

The fast-evolving AMG technology can potentially offer many benefits and applications across different domains. It can, for example, enable musicians and composers to explore new ideas and musical possibilities, adapt dynamically to coincide with on-screen action in games and visual media, or democratize music creation by helping individuals with different disabilities to create their own compositions.

However, despite such significant progress, it is clear that many challenges remain. As research continues, further advances can be expected in generation and evaluation techniques, along with improvements in the quality of generated music. The most common problems related to AMG include (Briot & Pachet, 2020):

- Control: Ensuring that the generated music conforms to specific tonality, rhythm, and other musical constraints can be difficult, especially for deep learning based models.
- Structure: Generated music often lacks a sense of direction or coherent structure, resulting in music without a clear narrative.

- **Creativity:** The balance between imitation and originality is crucial to avoid plagiarism and generate new music.
- **Interactivity:** Allowing human composers to influence the music generation process is critical to ensure the uptake of AMG systems and their integration in music creation tools.
- **Evaluation:** Assessing the quality of generated music is a complex task. As objective metrics are not able to capture all the intricacies of music, subjective evaluation methods, such as listening tests, are required to complement them. This challenge will be discussed in depth in the remainder of the paper.
- **Copyright:** The determination of copyright and ownership of machine-generated music is complex. In many cases AMG is used to produce public domain results but, when this is not the case, copyright ownership has to be clearly established (Frid et al., 2020).

This work aims to enhance the research and resources available for addressing the evaluation problem. As AI-generated content in the arts continues to advance, the demand for standardized and easily implementable evaluation methodologies grows. Furthermore, the potential use of these AI systems in education and model training highlights the need for robust safeguards to ensure the accuracy and reliability of generated content before it is disseminated more broadly. To address this, we propose a clear three-step process for developing and implementing a user-based evaluation methodology for automatic music generators, adaptable to researchers' specific needs.

When assessing AMG generators and the music they produce, a broad range of characteristics must be considered. Key evaluation criteria include the quality of the generated music, user acceptability, comprehensibility, structural coherence, emotional expressiveness, engagement, stressfulness, and applicability to specific use cases (although this can be expressed in a myriad of formulations). These aspects can be evaluated through objective methods, subjective assessments, or a combination of both.

1.1. Related works

Numerous studies have explored specific tools for music generation and evaluation, yet comprehensive and standardized approaches to assess computer-generated music remain scarce. Previous methodologies can be broadly categorized into objective and subjective evaluation frameworks, each with distinct limitations and potential applicability to current challenges.

- **Objective evaluation employs quantitative** metrics to measure the quality of generated music. Cosine similarity, for instance, evaluates the resemblance of generated sequences to human-composed ones (Chuan et al., 2020). Similarly, Kullback-Leibler (KL) divergence estimates the difference between probability distributions of AI-generated and human-composed music (Jiang et al., 2020). While these metrics are helpful for quantifying quality, they often

fail to capture subjective listening experiences. Other tools, such as Fr chet Audio Distance (FAD), assess perceptual quality using pre-trained deep learning models (Kilgour et al., 2019). FAD is particularly effective for general audio evaluation but requires substantial audio data for meaningful results. Additionally, symbolic music evaluation frameworks, such as heuristics based on music theory (Dervakos et al., 2021) and toolkits like MusicPy (Dong et al., 2020), focus on pitch, rhythm, harmony, and style. However, their relevance to non-symbolic music remains questionable due to their dependence on automatic transcription systems (Bittner et al., 2022). A further interesting approach, also for the symbolic music domain, is presented in Yang and Lerch (2020) that tries to create a combination of different objective metrics that provides results that try to be nearer to what human experts would expect of subjective evaluation results. This work also highlights the challenges of balancing subjective and objective approaches to achieve comprehensive assessments

- Subjective evaluation involves human listeners assessing AI-generated music based on their preferences and perceptions. This approach typically uses various survey-based tools, such as Likert scales (Chu et al., 2022) and others. A common test is the usually referenced as musical Turing tests (e.g., Hern ndez-Orallo (2020)) which tries to discriminate human-made music from AI-generated music. As Ariza (2009) comprehensively explains the musical Turing test is a misused term which should be replaced for "Musical Directive Toy Test" (MDtT) or "Musical Output Toy Test" (MOTt) depending on the scenario. These subjective evaluations can be used to evaluate parameters like creativity, structure, and emotional impact. While subjective evaluations provide invaluable insights into listener experiences, they often lack standardization and scalability. Hybrid approaches have also emerged, combining objective metrics with subjective user feedback. For example, Jordanous (2012) proposes a methodology to evaluate systems for creativity based on finding the most widely used concepts in creativity description (Social Interaction and Communication, Interaction and Emotional Involvement and Domain Competence...) and provide a weight for these components based on the importance given by evaluators and their musical experience. Later Four AMG improvisation generators are evaluated using a 21 point Likert scale for each component and a weighted average for each is provided.
- Mixed evaluation uses both objective evaluation metrics and user-based evaluation. As an example Sturm and Ben-Tal (2017) uses four different objective evaluation techniques to evaluate the transcriptions produced by a symbolic AMG system and also uses a very open expert-based evaluation in which a small set of musicians write, in free format, their opinions on the system output. Huang et al. (2021) utilized both quantitative metrics and user surveys to evaluate AI-generated music mash-ups. However, the integration of both methods remains clearly underexplored.

Another set of interesting works are theoretical frameworks that discuss mainly the theoretical aspects of artistic creativity such as Wiggins (2019). In this case it is applied to development of music from the 10th to the 20th century with a very formal approach. These approaches are beyond the current state of development of AMG.

While some works like Jordanous (2012) present evaluation methodologies most of the discussed works are not based on a methodological approach but take an ad hoc approach to evaluate specific music generators. Thus, despite the number of work that implement generated music evaluation, a standardized methodology for evaluating these systems is still lacking.

1.2. Human vs objective evaluation

A potential issue associated with evaluation studies relying on objective metrics is that, while a given system consistently yields identical results with identical inputs, it may not produce comparable outcomes with inputs that users perceive as analogous. This challenge frequently arises in art-related evaluation studies, where technically disparate inputs may exhibit a significant level of similarity grounded in shared cultural experiences, emotional responses, or additional factors.

Having a wide pool of users, an expert panel, or both to validate an AI music generator can provide a range of metrics that corroborate the generator's actual performance in a culturally and technically informed manner. In addition, they can also add an extra layer that guarantees the validity of the objective metrics that were used to train and validate the model.

On the other hand, evaluation studies based on human judgments are much more expensive and more difficult to scale up than those based on objective measurements, making them unsuitable for gathering data for model training (Yang & Lerch, 2020). Even though it may be interesting to be able to distinguish between human and computer-generated music using a musical output test, it has long been apparent that these types of evaluations have very little value in real-world scenarios (Chollet, 2019).

1.3. Aims and objectives

In this study, our main objective was to develop a methodology that can be adapted to the specific needs of validation and evaluation studies in the field of AI-generated music. As explained in subsection 1.2, there are several advantages to having human participants (experts or otherwise) subjectively evaluate art-related generation. Our intention was to provide a comprehensive set of tools that can help researchers perform user-centered evaluations that can be tailored in length and which focus on their own particular objectives and interests.

To the best of our knowledge, current research only covers specific evaluation tools or theoretical categorizations of strategies (Xiong et al., 2023), and this creates a situation in which designers of validation studies may need to read through several (and in many cases lengthy) articles before being able to create a tool-set suited to their intended evaluation.

Finally, through a proof of concept study, our intention is to showcase the application of the proposed methodology and, more importantly, the usefulness of several of the proposed tools for the validation, evaluation and comparison of AI music generators. We hope that this study will serve to advance awareness of several validation tools and contribute to the standardization of validation studies for generative AI in the arts.

This research aims to address the critical gaps in current evaluation practices, fostering a more standardized and inclusive approach to assessing AI-generated music. By aligning with best practices in user-centered design and computational creativity (Sturm and Ben-Tal, 2017; Hernández-Orallo, 2020), this study contributes to advancing the field and ensuring its practical relevance.

2. Materials and methods

We propose a methodology for tailoring evaluation studies to the specific needs of researchers. Figure 1 shows a general view of the method. Following a simple three-step process, researchers should first define their study by answering a series of questions with the help of the adaptability decision tree (see Figure 2). Secondly, they use the adaptability matrix (Figure 3) to choose the tools that are most appropriate for their study. Finally, they implement the selected range of tools following the guidelines and descriptions in the respective subsections in section 2 (Materials and Methods). If necessary, they can further familiarize themselves with the tools using the extended literature provided.

2.1. Evaluation study definition

As a prior step to any evaluation study, especially on AI generation systems, it is necessary to establish the purpose of the evaluation. A clear target for evaluation facilitates both the design and the implementation of the study. We identified two main categories of evaluation studies: those intended for model training, and those intended for confirmation and comparison of model performances.

The first category requires highly replicable results, minimal variability in outcome, and a very high number of measurements and iterations. The use of objective measures that do not rely on costly user studies, but on mathematical formulas or automatic training is therefore a must. In this case, scalability is the most important thing when training the model, and problems such as the sensitivity of the measuring tools are secondary.

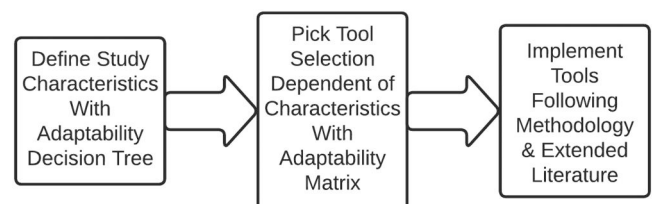


Figure 1. Meta-methodology overview.

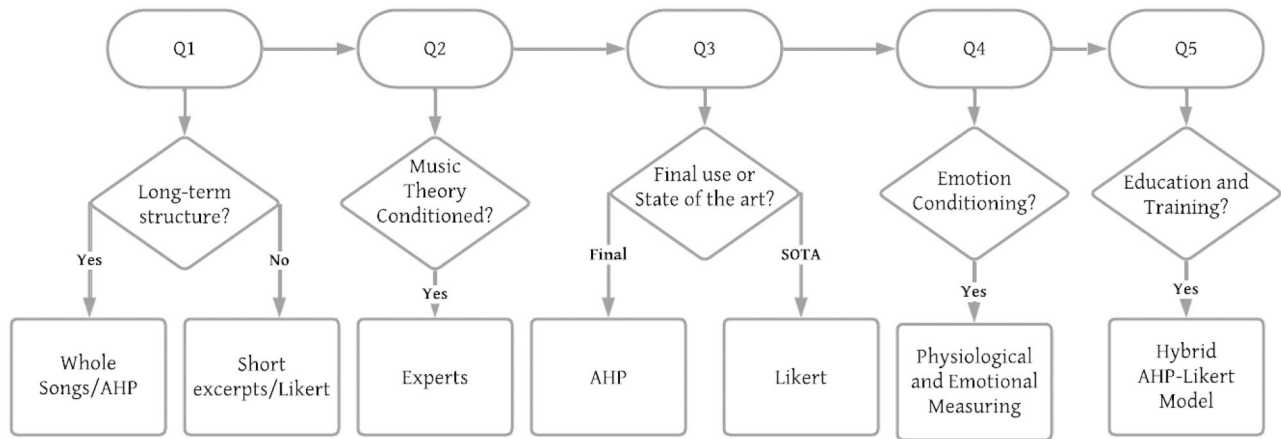


Figure 2. Adaptability decision tree. Q1. Is the generator intended to be used as a compositional tool or to produce finished pieces? Q2. Is the model conditioned? Is this conditioning dependent on music theory? Q3. Is the evaluation intended to establish the final usability of the model? Does it compare it to other state of the art systems? Q4. Is the generator intended to be able to tailor music to particular emotions? Q5. Is final quality control part of the evaluation? Is the model intended for educational purposes or for the training of other systems?.

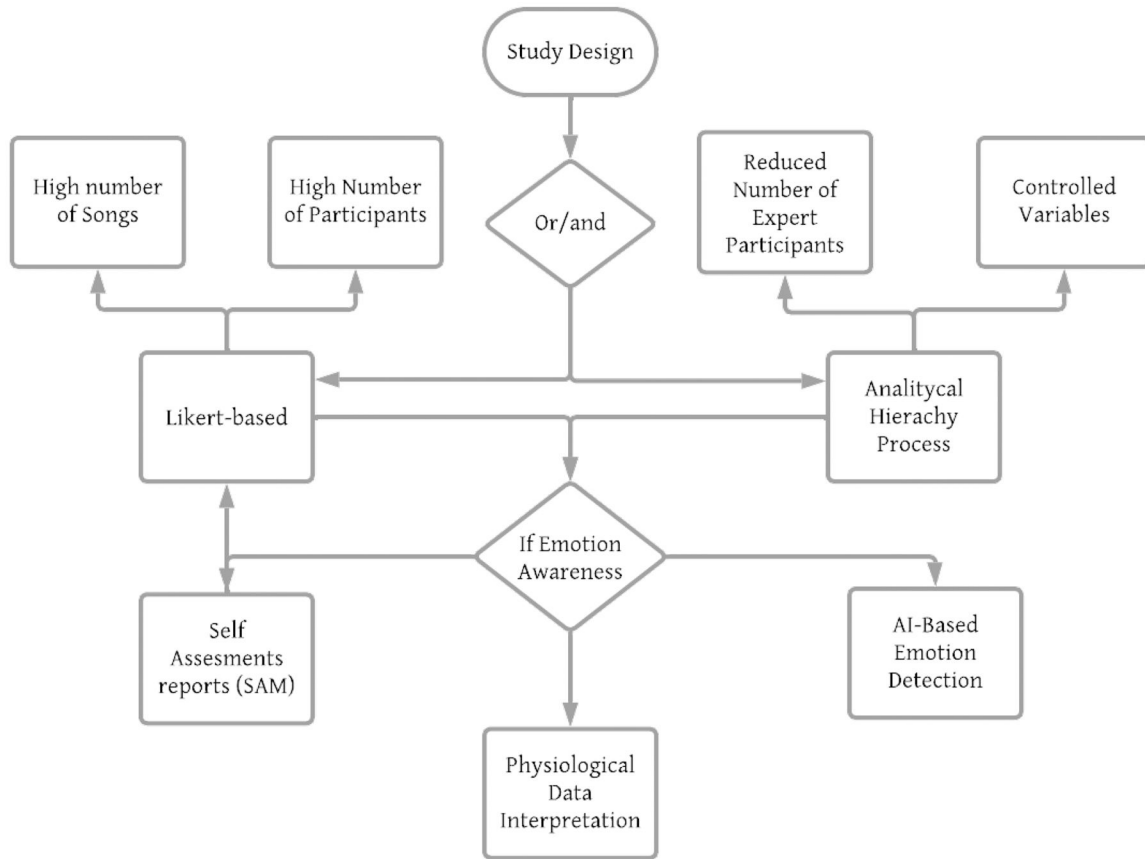


Figure 3. Adaptability matrix.

The second category of evaluation studies can be tackled from many directions. Studies like those of Chu et al. (2022) based on surveys with 9 parameters and using a 7 point Likert scale and Ferreira et al. (2023), based on a musical output test, show how a system's performance can be measured with many different tools with great results. In Wang et al. (2021) an objective metric for music generation is evaluated through subjective evaluation. The experiments they carried out demonstrate that objective evaluation is still less effective than subjective evaluation. This is a significant argument for continuing to develop better subjective evaluation methodologies.

In this article, we propose an adaptable model for designing human-based evaluation studies for the second category of the studies. This model provides the necessary tools and steps to design evaluation studies tailored to the specific needs of researchers and the music generator they are testing. Based on our previous research into the state of the art of AMG, in Civit et al. (2022), and taking into consideration several studies with different objectives and a variety of possible strategies, we designed a set of questions which, paired with a decision tree, can guide the designers of future evaluation studies in their tasks. Answering the questions in

Figure 2 can help researchers define their generator and its evaluation study, and thus determine the most appropriate tools.

The answers to the questions underline relevant characteristics that the study may require. In Q1, for example, researchers may need to use longer, complete pieces of music that require the judgment of long-term structures if they intend to generate finished pieces. This is not the case when using AI generation as a composition assistant tool, where only the quality of short excerpts of music may be relevant. Taking into account that any study made with users requires a considerable amount of time, studies related to longer pieces will benefit from the use of the analytical hierarchy process (AHP) methodology rather than the other proposed tools, because it requires more user time per piece of music, but provides significantly more data, with greater consistency, for a smaller sample pool (Saaty & Özdemir, 2014).

Q2 highlights the need for the participants to be experts in music if the model is conditioned by music theory rules. The main aspects of the different evaluation methods are shown in Figure 3. The use of AHP is recommended for expert judgement as it requires a smaller pool of participants. This is a common situation when sourcing experts for evaluation studies. In Cideron et al. (2024), participants sourced via the Amazon Mechanical Turk platform, were required to have experience listening to music (over 6 years), this type of requirements are common when sourcing experienced listeners, but are much harder to accomplish when looking for expert composers or musicians. If the use of expert judgment is not required, as is the case in any general preference study that does not require music-specific knowledge, the use of Likert-based surveys could be highly beneficial. Having more than 20 participants and a good sample pool can give good results faster and with much less effort on the part of the participants. In addition, tools such as Amazon Mechanical Turk and CROWDMOS (Ribeiro et al., 2011) can be used to recruit and select a very high number of participants, producing results that are easier to generalize.

The final usability of the generation model, referenced in Q3, requires a dual approach with regard to evaluation. Firstly, many aspects of the generator have to be taken into account, from performance characteristics such as music quality evaluation to human-computer interaction characteristics such as ease of use. Secondly, the performance of the model should be measured in real-world situations, i.e., the music it generates should be compared to human-made music. With these characteristics in mind, the pairwise comparison-based AHP was considered uniquely suitable for the task, as it excels at comparing two elements (human and AI generated songs) and can take into account many variables with different statistical weightings. However, if the purpose of the study is not for the generator to be publicly released, but to advance scientific knowledge, the comparison between multiple state of the art systems and their respective generations may be too time consuming for the AHP methodology, and the much faster and simpler to use Likert-based survey will be more suitable for the task. An example of this

approach can be seen in studies such as Meta (Copet et al., 2024), where a large user pool and a large sample pool from various generators are used for user evaluation with the aforementioned CROWDMOS system.

Many current commercial generators such as Frid et al. (2020) generate music based on emotions and music genre (e.g., happy techno or relaxing ambient music). Q4 is designed to identify whether the generator uses such features. For an affirmative answer, we propose a set of tools for self-assessment, physiological measurement, and emotion recognition that can be used with participants in the study (see subsection 2.7). These tools make it possible to measure and self-determine emotional states and should be used for both dataset tagging and study evaluations when the emotional state needs to be estimated reliably.

Q5 serves as a safeguard for future iterations of the model, future research and, in general, the passing on of misleading data to future generations. When the purpose of the model is to generate music for training other systems, or for use in some form of educational technology, there is a considerable risk that incorrect data may be passed into the future. This may have compounding effects (Osoba et al., 2017), such as misleading tagging in the dataset or generations, performance scores that do not coincide with real-world human perception. The use of the hybrid Likert-AHP methodology proposed in subsection 2.2 therefore becomes even more relevant, as the crossing of data from several different sources diminishes such risks.

2.2. General model and adaptability matrix

We propose an adaptability matrix (Figure 3) that allows researchers to design different evaluation studies depending on their resources and on the ultimate purpose of the study. This methodology is intended to minimize the risks of human error-induced biases and incorrect data.

After defining the evaluation study (see subsection 2.1), the choice of tools and methodologies to implement the study can be overwhelming. Whereas Figure 2, serves to define and delimit the evaluation study and can suggest some methodologies, Figure 3 shows the main tools proposed for designing a evaluation study and identifies the situations for which they are best suited.

Here, some important general characteristics need to be taken into account:

- The possible number of participants.
- The time that participants will need to complete the study.
- The need to include expert participants.
- The possibility of using a controlled environment with controlled variables.
- The number and duration of songs.
- The need to measure emotional states and the time that researchers may need to interpret the data.

As a general guideline for each of these topics (see Figure 3) we suggest:

- Using Likert-based surveys when more participants (usually 20 or more) are available, due to its ease of use and its consistency being tied to a minimum number of participants. The possibility of using online systems for recruitment and screening is an added bonus. On the other hand, AHP is better suited to smaller studies.
- AHP surveys are much more time consuming than their Likert-based counterparts and this needs to be taken into consideration when designing the study. Nevertheless, AHP can offer more detailed comparisons between alternatives.
- If expert participants are required AHP is recommended because, expert users are harder to source and it is therefore common to have a smaller reviewer pool.
- For emotional state measurement and real-time response tracking, controlled environments can improve the overall quality and reliability of a study. Tools like Noldus FaceReader and HRV measuring instruments work best when conditions, such as light or sitting position, are controlled. This makes them harder to use in combination with remote evaluation protocols. If implemented, such processes need to be carefully designed and supervised.
- For long-term structure analysis, songs with at least 30s duration should be analysed. Such longer formats that may take up more evaluators' time may benefit from a more thorough analysis using AHP methodology. In cases where many songs are presented, such as multi-genre generation, a faster Likert-based approach may be preferred.
- In [subsection 2.7](#) a range of tools are proposed for studies looking at generators with an emotion component, like those based on natural language prompts, which are able to produce "Happy folk song" type music.

All these topics are addressed in greater detail in the respective sections dedicated to each tool/methodology: Likert in [subsection 2.5](#), AHP in [subsection 2.4](#) and emotion measurement in [subsection 2.7](#).

2.3. Samples and participants

A major concern in any evaluation study is the selection of the samples and participants. When creating the sample pool, in our case pieces of music, the selection should take into account the objective of the study. Studies intended to evaluate generators which do not seek to produce any particular genre or are conditioned by theory-based criteria but merely, generate "music," may require a wide selection of musical pieces to be representative of a common generality. Music generators like DeepBach (Hadjeres et al., 2017) which are very style-specific, can be evaluated using a more representative but smaller sample pool when assessing specific criteria (for example, the quality of voice leading in a 4-part Bach-style harmony).

Another important aspect related to sample selection is commonality. To be easily compared, musical fragments should all have as many common characteristics as possible and be more or less of the same duration, regardless of whether they have lyrics. To avoid biases in the study, it is also very important for all the music in the sample pool to

have the same perceived loudness. Loudness should ideally be normalized to -14 LUFS as proposed in the guidelines of Katz (2015). This normalization in volume accounts for the difference in volume that different generators or musical sources may output. It facilitates any experimental design by allowing participants to compare different musical pieces by setting a base volume for all the experiment without the need to readjust the volume of the device while changing pieces. This very conservative loudness level allows comparison of different pieces without modifying the inner dynamics of the pieces, and thus preserving musical characteristics that may be linked to emotion or preference Schubert (2004).

Secondly, the participant selection must be addressed taking into account the idiosyncrasy of music, both as an art form (with its technical implications) and as a cultural experience (Agawu, 2006). With this in mind, the proficiency of participants in music can be considered both from a technical point of view and from the perspective of their exposure to music: the opinion of a wide array of informed, culturally experienced listeners may be as valuable for the researchers as that of expert musicians or composers. Furthermore, when considering the emotional implications of music and its impact on listeners, it is important to address the need to establish a neutral initial emotional state to avoid possible biases. Tools such as the Profile of Mood States (POMS) (Cayrou et al., 2000) can help verify this neutrality of mood and constitute the best option for initial testing.

2.4. Analytical hierarchy process and applicability

The analytic hierarchy process (AHP) is a multi-criteria decision making (MCDM) approach for structuring multiple choice criteria into a hierarchy, assessing the relative importance of those criteria, comparing alternatives for each criterion, and determining an overall ranking of the alternatives. It is particularly useful for teams working on complex problems that involve human perceptions and judgments.

AHP has been applied in several fields, including education and health, to address complex decision making and evaluation problems (Dolan, 2008). The process can be implemented with the aid of software tools like the AHP-OS online app (Goepel, 2018). AHP provides a systematic, structured approach to evaluating automatic music generators and human-composed music, or for comparing the two. Taking into account different criteria and their relative importance, and adjusting them depending on the context or in line with the preferences of the evaluators or an external group, it can be used to evaluate generators while considering many characteristics.

The process follows the steps shown in [Figure 4](#). A practical demonstration of how to implement AHP can also be found in [section 3](#) (Results). The described implementation is illustrated in [Figure 6](#).

First, it is necessary to define the criteria that are most relevant for evaluating music generators. These criteria could include musicality, creativity, originality, melody quality, harmony, rhythm, and the user-friendliness of the interface. Of the many possible criteria, particular attention should be paid to long-term structure, noise-to-signal ratio perception

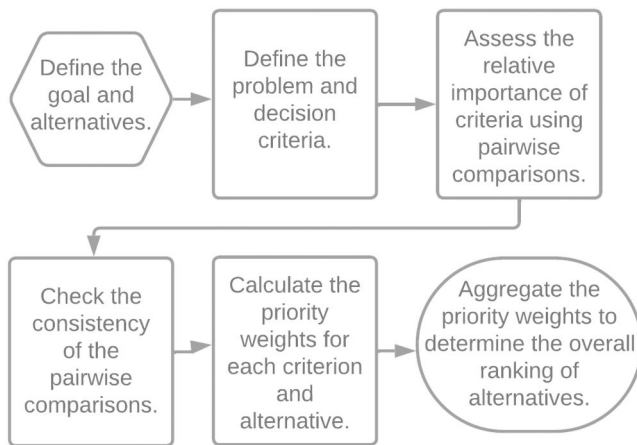


Figure 4. AHP steps.

(in particular for non-symbolic generators), and general user preference.

A second crucial step consists of organizing the criteria into a hierarchical structure. The top level is the overall objective (e.g., “Evaluate the effectiveness of automatic music generators”), and the next level specifies the main criteria (e.g., “musicality,” “creativity,” and “user-friendliness”). Sub-criteria can be added below the main criteria (under “musicality” sub-criteria like “melody quality” and “long-term structure” can be added).

For each pair of criteria at the same level, listeners are then asked to make pairwise comparisons and determine which criterion is more important. AHP typically uses a 1 to 9 scale in which odd values are significant steps (with 1 being of equal importance) and even values are intermediate steps. In this stage a comparison matrix is also created, based on the listeners’ responses. Tools such as AHP-OS provide an interface for listeners to register their comparison judgements and construct the associated comparison matrix. Although this step can be skipped if researchers prefer to use an equal criteria weight distribution, it is nevertheless one of the major advantages of the system, as it allows listeners’ criteria preference to be taken into consideration and it even allows the researcher to use two different groups (e.g., to use expert musicians’ judgements for the pairwise comparisons but with a criteria weight distribution established by a non-musician general population).

After completing the survey, the pairwise comparison results are used to calculate the weighted scores for each criterion. This is done by means of eigenvalue and eigenvector calculations on the comparison matrices carried out by the AHP software tool. The weights represent the relative importance of each criterion.

It is now possible to evaluate different generators using the criteria and their weights. The assigned scores reflect how well a generator performs on each criterion. Finally, an aggregation step is executed by multiplying the scores of each criterion by their respective weights and adding them to produce a final score for each music generator. This will generate a total score that represents the overall evaluation of the generator.

It should be noted that the AHP methodology and several AHP development tools such as AHP-OS (Goepel, 2018), include very thorough data diagnosis components, allowing for the evaluation of analysis results. Moreover, another very positive aspect of the AHP methodology is the possibility of dividing participants into clusters with high consensus among themselves, but with low consensus when compared to other clusters. This can be done thanks to the invention of a consensus indicator (S^*) that shows how much the members of a cluster agree with each other. This feature can be used to divide listeners into coherent sub-groups based on their opinions of the different alternatives, and, as such, is potentially very useful for validating generators that may include genre transfer as part of their features, where the possible biases on the part of listeners towards particular musical genres can be decisive in the evaluation. Another benefit of this clustering methodology is that it can easily detect outliers, whose opinions may be beneficial to take into consideration, whereas other methodologies may simply discard them.

When assessing a group with a wide-ranging opinion pool, such as listeners with different musical expertise, and outlining their expectations and priorities for human or AI generated music, it is common for the collected data to indicate a good logical coherence but for the AHP group consensus indicator (S^*) not to reach the desired level, as there may be contradictory judgements. To address this issue, AHP implements a consensus indicator and a method for distributing the group into subclusters. This usually leads to a small subset of coherent groups with much higher consensus. The differences between these groups are always worth analyzing and should be justified in the study.

For our proposed methodology, AHP should be prioritized over Likert-based surveys in studies with small user pools (under 20 users) or small sample pools (under or around 8). As AHP functions well in small group settings due to its high number of cross-checks between different criteria and its built-in consistency metric, it works best in the aforementioned scenarios. AHP should also be given preference or used in conjunction with Likert-based surveys in those evaluation studies where the objective is to rank different generator systems against each other. The user-weighted multiple criteria of AHP studies allow researchers to compare generators using a multilayered approach that can result in very thorough insights that go far beyond general user preference.

2.5. Likert survey and applicability

The Likert scale (Eerola et al., 2018) is a rating system widely used in surveys to estimate opinions or perceptions. Subjects can choose from a set of responses (usually 5 or 7) to a question. Table 1 shows some typical Likert scales. These scales are widely used in both social and educational research and are the basis of many software evaluation studies centered around user evaluation. When using Likert scales, the researcher must consider issues such as response categories (the values in the scale), the size of the scale, the

Table 1. Examples of likert scales.

	5	4	3	2	1
Agreement	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
Frequency	Always	Often	Sometimes	Rarely	Never
Importance	Very Important	Important	Moderately Important	Slightly Important	Unimportant
Likelihood	Almost Always True	Usually True	Occasionally True	Usually Not True	Rarely True
Quality	Excellent	Good	Fair	Poor	Bad
Distortion	Imperceptible	Just perceptible not annoying	Perceptible slightly annoying	Annoying not objectionable	Very annoying objectionable

direction of the scale, the ordinal nature of Likert-derived data, and the appropriate statistical analysis of those data.

Likert-derived data are usually treated statistically as interval-level data (i.e., values on the scale have directionality and are equidistant). This allows the use of parametric tests like analysis of variance (ANOVA) or Pearson's product-moment correlation (Sullivan & Artino, 2013).

Of the tools that are available for designing Likert-based surveys, CrowdMOS (Ribeiro et al., 2011) has become widely used in the field of audio quality evaluation, and, by extension, music software evaluation. An open-source project that can be easily deployed on a web server, it offers a way to crowd-source Mean Opinion Score (MOS) studies for subjective evaluation of audio, video, and image quality. CrowdMOS has been successfully used in several music generation studies (Copet et al., 2024) and provides a convenient and cost-effective tool to perform subjective quality evaluations, using platforms such as Amazon Mechanical Turk to recruit Internet users to participate in MOS-like listening studies. It makes it possible to filter noisy annotations and outliers by removing participants who do not listen to the full recordings or who rate references (in our example, human-composed music) with scores below 85. It also provides a set of recommended strategies (Ribeiro et al., 2011) for cleaning data. The Likert scales for quality and distortion shown in Table 1 are widely used in CrowdMOS evaluations.

It should be noted that among the many advantages of Likert-based surveys, speed and ease of application are of the utmost importance. These characteristics make such surveys well suited to studies with many samples because the surveys themselves do not take up a significant amount of time from the participants, unlike methods such as AHP, described earlier.

For our proposed methodology, music generator evaluation studies that are not centered on the generation of a single specific genre, but may cover multiple music genres, should use Likert surveys in preference to AHP. Due to the wide variety of music sources involved, these types of studies require many more samples for a generator to be properly tested. The speed of the Likert application, in particular when users are making individual choices, makes it possible to evaluate a bigger sample pool of musical fragments. In our experience, fragment duration should usually be kept at around 30 seconds, as the best way to balance test speed with the possibility of implementing emotional measurements. This is, however, only an approximate time frame, and further research is needed to create more definitive guidelines. The use of the Likert based evaluation in this methodology can be implemented to accelerate the evaluation process when many pieces are to be evaluated. As an

“accelerated” method, a smaller amount of criteria for evaluation may be preferred, which may leave behind concepts such as long-term structure, motivic development or continuous monitoring of emotional state through a piece (which would be replaced for dominant emotion classification/recognition).

The possibility of subsequently cross-referencing a sub-sample pool of Likert data with an AHP survey should also be considered, in particular in sensitive studies dealing with generators for education or model training, where failure in evaluation could produce cascading effects.

2.6. AHP, likert, and the adaptable hybrid system

Both Likert-based surveys and AHP-based studies have significant advantages and drawbacks. Ponsiglione et al. (2022) proposes a hybrid methodology to combine both approaches. Figure 3 illustrates the possibility of integration of Likert and AHP based assessments.

In general, the combination of the two approaches offers a number of benefits. The use of two different data collection methods can, for example, help reduce errors and improve the accuracy of results. Complementary methods can also provide a more complete picture of the situation and can provide useful insights on the strengths and weaknesses of the studied generators that can help on their future improvements: AHP can, in general, provide guidance in more complex topics based on expert-user evaluation while Likert can benefit from a wider user pool with less expertise for simpler issues.

AHP and Likert based surveys can interact in different ways. As a first option AHP can be used to weight the Likert scale responses. In Song and Kang (2016), a method is developed where Likert scale surveys are used to reduce the number of pairwise comparisons. Ing (2021) takes a very practical approach, using differences in mean data from a previously available Likert scale survey to produce answers for the pairwise comparisons required by AHP.

Another possible integration can use Likert scale surveys to validate AHP results (Gutknecht et al., 2018). Likert scales can be used to compare the relative importance of the criteria or alternatives obtained from both methods. In this approach, AHP is used to prioritize criteria or alternatives, while Likert scale surveys are used to directly gather user feedback on each criterion or alternative.

A variation of this implementation can provide expert supervision of music dataset tagging in music generation. Comparing the results from the larger and non-expert based Likert surveys to the shorter AHP questionnaires that use expert, reaserch can detect results that have a wide variation

on assesment ranges between different groups. This can indicate an error either by experts being outside their particular realm of expertise (as could be the case highlighted in subsection 4.2) or the criteria being too technical to be properly assessed by the untrained group. Rooting out errors, in particular in the dataset gathering stages is a very important topic that can highly impact music generation, thus requiring further and specific research on the topic.

As a third option, AHP and Likert scale surveys can be used to create a dynamic, interactive decision-making process. For example, a web-based tool could be developed that allows stakeholders to input their responses to the Likert scale for different criteria. The tool could then use AHP (e.g., in the way proposed by Ing (2021)) to generate a prioritized list of alternatives. The experts could then review the results and provide feedback, and this feedback could then be used to refine the AHP model and generate a new prioritized list of alternatives. This process could be repeated until a satisfactory solution is achieved.

Finally, user opinions could be collected using Likert scales, which using a methodology similar to that proposed in Ing (2021), could be used to create inputs for an AHP-based analysis. Alternatively, users could directly answer AHP pairwise comparisons and then use a Likert survey as confirmation and complementary information. An approach combining direct pairwise comparisons with “virtual” comparisons obtained from a Likert survey could also be implemented and analyzed.

To further corroborate evaluation results, including also emotion / engagement / stress analysis might be very useful as a means of obtaining actual user feedback in real time.

2.7. Emotional state measuring

Measuring emotional states in listeners can be extremely useful as a research tool, particularly in creative fields. Having listeners self-evaluate their emotional state using a survey based on self-assessment manikins (see Figure 5), conducted after having listened to each piece of music, is a very simple way to integrate emotional measurements into a evaluation study. Self-assessment manikins (SAM) are a

picture-oriented approach to measuring emotional responses that have been used successfully in studies related to numerous topics, including music education (Cuadrado et al., 2020). SAMs aim to estimate three central features of emotional responses: perceived valence, perceived arousal, and perceptions of dominance. As a (usually) 9 point scale, they constitute a tool that is easy to integrate into a fast-paced Likert-survey model. Researchers should consider that other Likert-based surveys for emotional estimation exist and been used to asses AI vs human generation (Diwanji et al., 2025) but SAMs are a widely standardized tool.

To monitor emotional states in real time by means of a faster, more objective, emotional estimation, a complementary study based on the movement of face muscles or the measurement of physiological signals using wearable devices can be implemented. In this way, the researcher can compare and assess the validity of the listeners’ self-assessments.

The most widely used physiological signals are parameters related to heart rate variability, such as interbeat intervals (IBI), and galvanic skin responses (GSR) parameters related to sweat gland activity. Many studies use the certified medical device Empatica E4 wristband (McCarthy et al., 2016) or its current successor Empatica EmbracePlus. A significant number of studies have used the Apple Watch (Hernando et al., 2018; Hirten et al., 2021) for HRV studies, although it lacks any medical device certification. This device and other common smartwatches are a good alternative that can help reduce study costs and improve test scalability. Quality measurements of HRV are very important in regards to emotion estimation, as some devices may not provide the best accuracy when being devised for sport tracking. Nevertheless, there is a significant evidence of the accuracy of some commercial sport trackers when compared to medically certified equipment (Lui et al., 2022), thus providing grounds for their use on other HRV measurement related tasks.

The recordings of wearable devices can be analyzed using statistical parameters calculated with tools such as the Kubios Heart Rate Variability (HRV) application (Tarvainen et al., 2014) and the MIT Medialab EDA explorer (Taylor

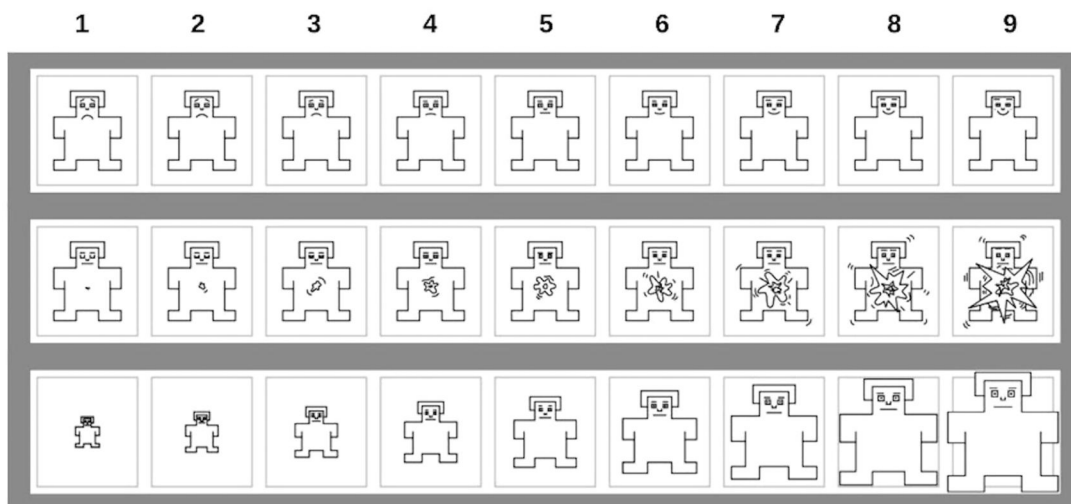


Figure 5. 9 point SAM scale (Knejzlíková et al., 2021).

et al., 2015) for the analysis of GSR. A comprehensive review of the works that use HR analysis to estimate emotions can be found in Ismail et al. (2024). A tool for emotion classification including a web-based interface can also be found in Bugnon et al. (2020). There are also several deep learning based tools suitable for this analysis (Muñoz-Saavedra et al., 2023). Directly measuring physiological signals usually requires a significant effort to process the data if good quality emotional information from listeners. Most researchers consider that at least 30 seconds of stimulus recording is necessary to be able to successfully carry out HRV analysis (Tanoue et al., 2023). However, it has been suggested that much shorter time analyses are suitable for valence prediction (Schippers et al., 2018). An implementation of these short-timed stimuli is further explored in section 3 (proof of concept). Nevertheless, the disadvantage of having to use longer durations to estimate general emotional states, as well as the preference for using these devices in controlled or semi-controlled environments to avoid confounding parameters may hinder the measuring of subtle emotional states that might briefly arise in music and that can also be shaped by the context of the listener. Improving these measuring technologies can improve a deeper understanding of the emotion to music detection in evaluation studies.

A consideration has to be made in regards to the relations between measured signals and emotions. The physiological responses that different individuals may have to emotions can vary significantly, adding difficulty to the development of universally applicable models (Ahmad & Khan, 2022). As the process of labeling physiological data with emotional states can be subjective, this can impact the reliability of the training data. Furthermore, Physiological signals are often noisy and contaminated with artifacts, requiring careful pre-processing to extract reliable features for emotion recognition. Different physiological signals may require different pre-processing techniques, which adds complexity to the development of robust emotion recognition systems.

As a second option, facial expression-based emotion detection systems can be implemented to further assess the emotional state of listeners. Most of these systems are based on the facial action coding system (FACS) (Rosenberg & Ekman, 2020), which encodes individual facial muscle movements based on visual observation of changes in facial expression. This system is widely used for both expression analysis and facial expression synthesis. FACS is based on the analysis of “action units” (AU) related to specific movements in areas of the face. FACS can be used to determine listener emotions in real time, as these are directly related to the activation of specific AUs (Kodra et al., 2013).

FACS can be coded manually, but this process is time-consuming and requires very well trained experts. Currently, almost all experiments use automatic action unit recognition. Programs like the open-source OpenFace (Cuculo & D’Amelio, 2019), Noldus FaceReader (Skiendziel et al., 2019), and Affectiva iMotions (Kulke et al., 2020) can automatically detect individual AU activations in real time.

FaceReader and iMotion both also provide real-time probabilities for seven basic emotions, valence, and arousal. The ease of use of these applications, together with the unobtrusive implementation in evaluation studies (they mostly just require a face-cam capable of recording listeners while they are doing the test), makes face-based emotion detection probably the best alternative for easily obtaining emotional data during music evaluation and, in this way, complement and contrast listeners’ self-assessments. However, as with the previous tool, the length of the stimuli is an important consideration: It is important to compensate for the traces of positive and negative expressions on neutral faces by normalizing the emotion values for individual participants as proposed in Kayser et al. (2022).

3. Pilot study

As a proof of concept, we decided to evaluate the real-world performance of the Meta generator (Copet et al., 2024) against human-made music from the MUSICAPS (Agostinelli et al., 2023) dataset which was used for its training. We then compared with the Perceiver Music Transformer (Hawthorne et al., 2019), a symbolic generator validated against the MAESTRO Midi dataset (Hawthorne et al., 2019) of human-composed and interpreted music.

To evaluate the feasibility of our proposed methodology for developing music validation strategies, a preliminary experiment was conducted that incorporated the majority of tools referenced in Section 2. The study employed a counterbalanced design with four participants who examined 40 short excerpts of music. The four participant were all post-graduate students, two male and two female with ages ranging from 25 to 33 years old. All had backgrounds in music research with extensive listening experience, but did not self-identify as music experts or professional musicians.

The music excerpts were equally split between those generated by the specified algorithms and those sourced from human-composed music, with all pieces adjusted to a uniform loudness of -14 LUFS (Katz & Katz, 2003). Given the limited number of participants and the objective to compare the two generative models against human-composed music across multiple dimensions, as outlined in the recommended guideline (see Section 2), the Analytic Hierarchy Process (AHP) was selected as the principal validation instrument. Due to the generators not being designed as genre-specific, it was imperative to utilize a diverse sample pool. The broad scope of this sample pool necessitated significant user effort to evaluate the generators via AHP, thus AHP was applied only to a subset of eight pieces (two of each generator and two of each human dataset), while a simpler Likert-based survey was employed for the entire 40 music sample set, in accordance with the guidelines. All music excerpt were previously selected randomly among the produce of the generators and the pieces in the datasets.

All the listening test and ratings were conducted in a controlled listening environment. All listeners used closed-back headphones. This space was a quiet and evenly illuminated room. Participants were asked to listen to the 4 AHP

selected pieces in a random order and then complete the AHP assessment. Then they assessed all pieces in the Likert subset by listening and assessing one piece at a time in a random order. All four participants completed the study, both the AHP and Likert assessments.

Facial expressions and physiological data (HRV) were collected while listening to each piece for the first time, before the AHP survey.

The incorporation of these various parameters for a comprehensive methodological assessment resulted in each participant requiring approximately two hours to complete the experience. This duration can be greatly optimized in specific experiments (Civit et al., 2024) by refining the data collection process using the Adaptability Matrix (Figure 3).

Two AHP surveys were used to rate several aspects of the pieces. They were created using the design shown in Figure 6, and followed a two-tier criteria approach with three main criteria being compared for each song and including three subcriteria for each of these criteria. This made it possible to establish comparisons and rankings between all four songs in each survey for all of those aspects.

As can be seen in Table 2, human-composed songs 1 and 2 were clearly preferred by listeners in almost all aspects. The table compares human-composed to AI-generated symbolic piano music, and includes a weighting distribution compensated by the judgement of 4 nonexperts users in music. This

provides extra feedback on the pieces and generators, while also highlighting the possibility and usefulness of using two groups with different characteristics for the evaluation and weight distribution stages (note that this is not done in this pilot study). Table 3 compares AI-generated and human-composed audio-based music. It also shows how listeners rated human-composed music higher in almost all aspects. However, listeners rated song 7 (AI) as very close to human song 5 in its overall evaluation. This demonstrates the viability of generators if enough cherry-picking is done with the generated output results. It is clear that grouping all human-composed or AI-generated songs would provide less insight into the real capabilities of generated music, as listeners clearly rated some songs above others in most aspects. This illustrates the fact that it is still widely desirable to generate several songs and let humans choose among them.

As AHP surveys are time-consuming, only eight pieces of music were evaluated with that method, the others being evaluated in a much more direct Likert-based survey seeking the “more liked” (the highest valued) subcriterion using a 9 point scale ranging from 1 (I didn’t like it at all) to 9 (I loved it). The first objective was to establish whether the participants had a clear perceived preference between human-composed music and AI-generated music, and between symbolic music and raw audio music. To answer this question, an ANOVA (analysis of variance) was carried out on the general sample of

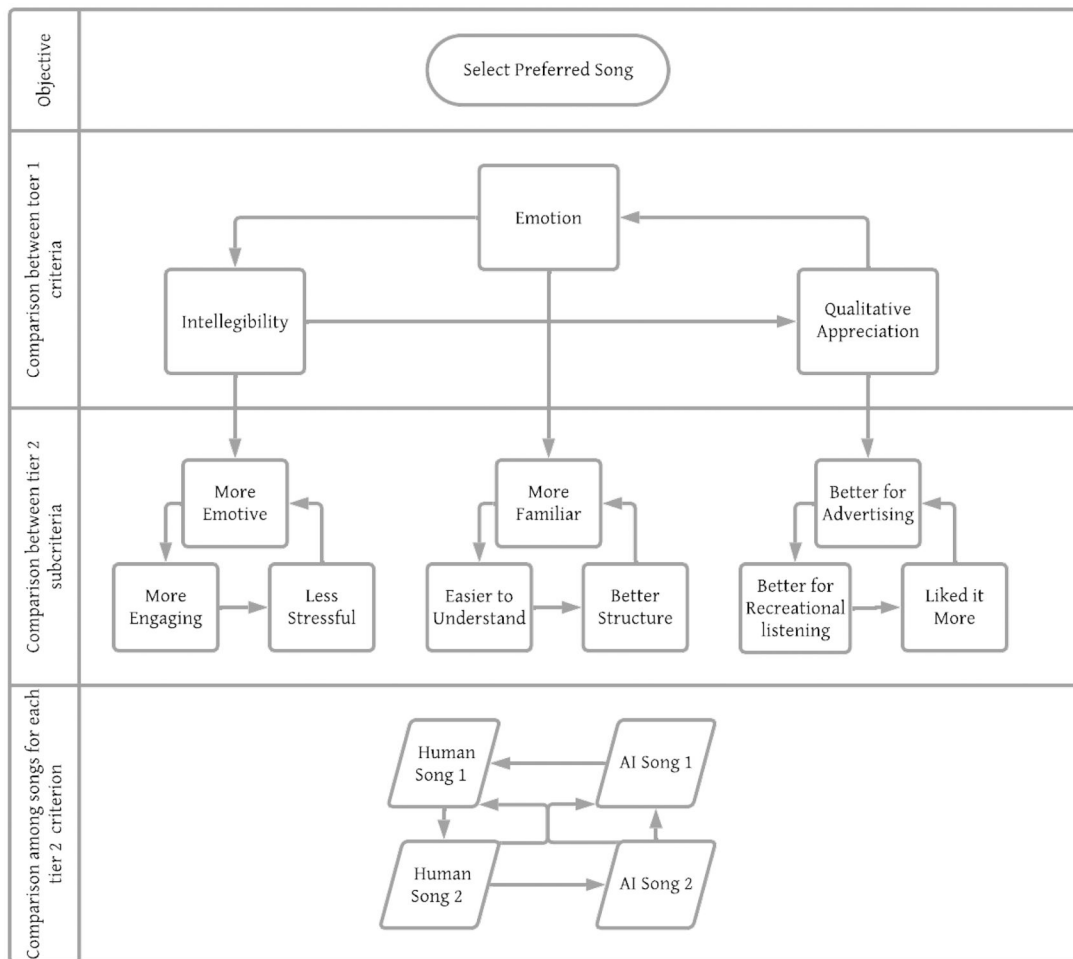


Figure 6. AHP study model.

Table 2. Symbolic music AHP expert evaluation with non-expert weight distribution.

				Human Song 1	Human Song 2	AI Song 3	AI Song 4
AI & Human Music	Emotion 0.455	More emotive 0.414	18.90%	0.495	0.296	0.09	0.119
		More engaging 0.479	21.80%	0.414	0.438	0.066	0.081
		Less stressful 0.107	4.90%	0.525	0.267	0.077	0.131
		More familiar 0.259	2.70%	0.44	0.425	0.062	0.073
	Intelligibility 0.106	Easier to understand 0.165	1.70%	0.443	0.406	0.074	0.077
		Better structure 0.576	6.10%	0.396	0.415	0.12	0.069
		Better for video advertising 0.092	4.00%	0.4	0.459	0.079	0.062
	Qualitative 0.439	Better for recreational listening 0.238	10.40%	0.407	0.429	0.1	0.064
		Liked more 0.670	29.40%	0.393	0.405	0.139	0.063
			100%	42.70%	39.02%	9.98%	8.17%

The numbers in the first column reflect the importance (weight) placed by the users on each of the main criteria as a fraction, i.e., emotion 0.455 means that the importance placed on emotions is slightly over 45% while the importance placed on intelligibility is around 10%. The third column tells as the importance of each subcriteria inside the criteria, i.e., More emotive has a weight inside the emotion criteria if about 41% while less stressful is valued only around 11%. The fourth column is the global weight of the subcriteria expressed as a percentage i.e., the product of the weight of the criteria times the weight of the subcriteria (e.g., $18.9 = 0.455 \times 0.414 \times 100$). The values in the remaining columns are calculated by the AHP algorithm considering the weight in the third columns and the answers provided by the listeners for the pairwise comparison among songs. These values reflect the grade given for the considered criteria for the particular song. The total sum of the grades for the different subcriteria for a song gives the global grade of the song. The addition of the global grades, if these are expressed as a percentage, should add to 100%. In our case we can see that song 1 gets the best grade (42.7%) while song 4 gets the worst grade (8.17%).

Table 3. Audio-based music AHP expert evaluation with non-expert weight distribution subcriteria weights are the same as the objective of the evaluation is the same.

				Song 5 Human	Song 6 Human	Song 7 AI	Song 8 AI
AI & Human Music	Emotion 0.455	More emotive 0.414	18.90%	0.311	0.467	0.133	0.09
		More engaging 0.479	21.80%	0.268	0.469	0.173	0.09
		Less stressful 0.107	4.90%	0.327	0.364	0.264	0.046
		More familiar 0.259	2.70%	0.289	0.39	0.262	0.059
	Intelligibility 0.106	Easier to understand 0.165	1.70%	0.246	0.372	0.326	0.056
		Better structure 0.576	6.10%	0.375	0.341	0.243	0.04
		Better for video advertising 0.092	4.00%	0.257	0.339	0.355	0.05
	Qualitative 0.439	Better for recreational listening 0.238	10.40%	0.304	0.374	0.28	0.043
		Liked more 0.670	29.40%	0.262	0.365	0.299	0.075
			100%	28.70%	40.58%	23.45%	7.24%

In this case (audio-based songs) The best song is also human composed and the worst is also AI composed but there are two songs (5 and 6) which have relatively similar grades. This shows that by selecting specific AI based songs we can get results that are similar to the not preferred human-composed songs. Criteria weight distribution could also be flattened to not account for user preference giving each subcriterion a weight of 11.11%.

generated and composed music. A two-way ANOVA model was first tested to ascertain whether the users' acceptance of different songs was directly linked to whether or not the music was human-composed and whether it was symbolic or raw audio. In this analysis, we obtained F values above 32 for human generation and above 14 for non-symbolic generation. In both cases p was below .0001, and therefore the null hypothesis had to be rejected. A Tukey test was performed for the two-way model, establishing a mean "like" difference between composed and generated pieces of 1.95 with a 95% confidence interval [2.62, 1.27] ($p < .0001$). The results for non-symbolic and symbolic music showed a mean "like" difference of 1.30 with a 95% confidence interval [0.63, 1.97] ($p < .0001$). Table 4 shows the mean values and the 95% confidence intervals for the different song groups.

Our experiment follows a within-subjects design in which all participants evaluate both types of music: AI-generated and human-composed. This design controls for individual differences by having the same participants evaluate both conditions.

To further test the songs, we implemented a set of tools to analyze emotional states. We used both HRV and GSR measurements and the FACS analyzer described in Subsection 2.7. These tools were paired with the SAM-based measurement done together with the Likert-based surveys, and their results

provide further insight into the emotional state of the participants while listening to the different pieces.

We also recorded the facial expression of the users while listening to the selected symbolic songs (1 and 2 human-composed, 3 and 4 AI generated). These data were analyzed using Noldus Face Reader and the corrections proposed in Kayser et al. (2022) were applied. In Table 5 we present the mean basic emotions with the associated valence and arousal values detected by Noldus Face Reader for the different songs.

In Table 6 we provide the mean and standard deviation of the valence and arousal value per song. In this table we can observe that Songs 1 and 2 exhibit high variability in valence, indicating diverse emotional perceptions among participants. This could be attributed to the complexity or ambiguity of these songs. Despite mixed valence ratings, Song 2 shows no variability in arousal, suggesting that it is universally engaging and has consistent emotional intensity across participants. Song 3, with its low valence and high arousal, appears to evoke strong reactions, potentially perceived as more unsettling or intense, likely evoking negative emotions such as anger or surprise. On the other hand, Song 4 is characterized by the highest mean valence and relatively lower arousal, suggesting that it is perceived as the most pleasant but with less emotional intensity.

Table 4. Grouped song acceptance mean rates and confidence intervals based on likert survey for all participants.

generated	audio	Lower	Mean	Upper
Human	Midi	5.1	5.68	6.25
Human	Audio	6.2	6.85	7.5
Generated	Midi	2.86	3.42	3.97
Generated	Audio	4.06	4.92	5.77

Table 5. Face Reader results main detected emotion values for valence and arousal for each song and participant.

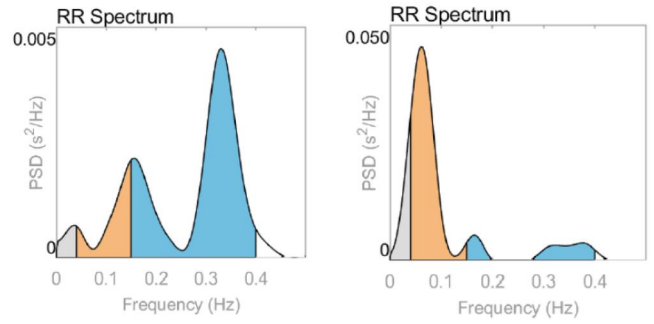
Song	Part.	FR_EMO	FR_VAL	FR_ARO
1	4	Angry(6)	3	9
2	4	Happy(8)	8	6
3	4	Surprise(6)	5	5
4	4	Surprise(3)	5	3
1	5	Angry(2)	3	3
2	5	Scared(7)	3	6
3	5	Angry(6)	2	8
4	5	Happy(8)	8	7
1	6	Sad(4)	3	5
2	6	Disgust(4)	4	6
3	6	Surprise(5)	3	5
4	6	Scared(5)	4	4
1	7	Happy(7)	7	7
2	7	Disgust(3)	3	6
3	7	Sad(4)	4	7
4	7	Sad(6)	4	6

Songs 1 and 2 are human-composed while song 3 and 4 are AI-generated. All songs are symbolic.

Table 6. Mean and standard deviation for valence and arousal per song for all participants.

Song	Mean Valence	SD Valence	Mean Arousal	SD Arousal
1	4.00	2.00	6.00	2.58
2	4.50	2.38	6.00	0.00
3	3.50	1.29	6.25	1.50
4	5.25	1.89	5.00	1.83

To measure HRV (heart rate variability), participants wore an Empatica E4 medically certified device during the tests. This device captures heart interbeat intervals, galvanic skin response, body temperature, and three-axis wrist accelerometer data (McCarthy et al., 2016). It has been shown that there is a clear link between the lower frequencies in the RR spectrum and the excitation capabilities of music (Dimitriev et al., 2022). An example of those effects is shown in Figure 7 where the left-hand plot represents the spectrum of a song that the participant clearly considered surprising. The right-hand plot represents the data for a song clearly considered happy. These analyses can be automated using machine learning techniques such as those applied in Muñoz-Saavedra et al. (2023). The validity of these data in our proof of concept is limited because we used the original short music samples to extract the HRV characteristics that are then correlated to valence values. The sample window must cover at least 30s as established in Schippers et al. (2018) and therefore we had to include some data after the listening period. Future studies aiming to use this technology should be careful to take the duration of the sample into consideration when designing the experiment.

**Figure 7.** HRV data for surprise (song 3) and happy (song 2) example of how emotions can be represented in HRV RR spectrum.

4. Discussion

4.1. Implications of the pilot study

The purpose of this pilot study was to highlight the viability of the methodology and tools presented in Section 2. Although the study had a quite small participant pool, the results are nevertheless interesting and may point towards areas that require future research.

First, in complete contrast to our initial hypothesis, the reviewers rated music generated in audio format as being much closer to its human counterpart than midi-generated music (see Tables 3 and 2). We initially assumed that with all midi recordings being very accurately normalized classical piano solos played by the same high-quality midi instrument (Analog Lab V), they would have ratings very similar to human midi music. Compared to audio-generated pieces, it seems that the generator's higher degree of freedom (apart from not using "human" voices, it has no other constraints) made it more interesting and more able to convey emotions (one of the highest valued criterion in the AHP survey). Further research is needed into audio vs midi generation, particularly with regard to effects on emotions, as this is a new and increasingly interesting field.

The second important implication of the pilot study is that although AI music generation is rated as acceptable (especially in the audio realm), the results are below those reported in the original Meta MusicGen study (Copet et al., 2024). Even though our study is too small to draw final conclusions, it highlights the need to independently validate the most used and ubiquitous music generators that may later be used as engines for music devices, music teaching tools, etc., or as a key elements in the short-term democratization of the music making process.

4.2. Music and culture for validation studies

As an art form that has accompanied humankind since the very beginning of our history, music has developed together with our culture, language, and forms of expression. Such a complex concept is usually the subject of whole dissertations, from historiographical research into its origin and development to cultural studies that build bridges between different interpretations. With many genres and subgenres, music is never just "music," and not considering this can pose serious problems for the future of our culture,

especially when dealing with AI-generated music. From its inception, our proposed methodology is intended to be genre-blind, capable of being implemented regardless of the genre of the music that is evaluated. Nevertheless, it is important to note that our research group is of western origin, and thus can be inherently biased towards favorising characteristics that are predominant in western music. Consequently, we advise researchers that are interested in non-western music genres to source participants and, particularly experts, that are well versed in these particular genres, and to consider if any of the steps in the application of the general model (Subsection 2.2) can be fine-tuned.

As explained in Civit et al. (2022), many current AI music generators are designed without considering that, depending on the corpus of music used for their training, they will have an implicit bias towards a specific genre, usually western classical music or western pop music (Moysis et al., 2023). A further problem arises when the generated product of such generators is used to train new models – something which to the best of our knowledge is not currently happening, but is nevertheless a more than likely scenario as generators improve in performance and datasets of AI generated music become available (Civit et al., 2024). In this case, those cultural biases will be passed on from generator to generator, and possibly even from generators to humans if we start using AI music generation for educational purposes. With such issues in the spotlight of debate, it is now necessary for validation studies to explicitly or implicitly address the genre specificity of music generators.

As an example, in our proof-of-concept study, we found that the MusicCaps dataset used to train the Google MusicLM generator (Agostinelli et al., 2023) has some inconsistencies in its tagging, having tagged a fragment of Jose Hidalgo's "*Las bragas que te compré*," a sevillana-style flamenco song, as a traditional Mexican folklore song, probably due to its lyrics in Spanish. Even though out of the particular scope of the proposed methodology, implementing the combined AHP-Likert method proposed in subsection 2.6 could greatly diminish the probability of these errors appearing, by cross-referencing user assessments with expert judgments. Validation studies should therefore check for consistency with the technical and cultural attributes of the intended genre specificity, and future research is needed to develop standardized tests able to validate the similarity between the intended genre and the output of a generator, especially in text-to-music models.

5. Conclusions and future work

The methodology proposed in this study is a step towards a general model for user-based AI music generation evaluation studies. In many cases, automatic quantitative evaluation methods are still insufficient for evaluating the quality of AI-generated works (Wang et al., 2021), so there is a clear need to create and improve standardized user-based evaluation protocols. Utilizing the suggested techniques allows researchers to considerably shorten the time needed to design a new study. They are also provided with a set of tools for obtaining

significant data that can be easily cross-checked with different variables and other measuring tools. The proof of concept presented in the study demonstrates the viability of the methodology and of the different tools proposed.

The standardization of a methodology for user-based evaluation studies aimed specifically at generative AI should be a major concern for the scientific community. The implications of this rapidly evolving technology are far-reaching and require studies capable of corroborating the effectiveness of AI generators and ethically guiding and enhancing them. Standardization will improve the ability of researchers to compare and contrast different systems in independently created studies. Moreover, this fast evaluation can guide researchers in improving their learning algorithms by fine-tuning areas that human evaluators may consider important and automatic evaluation tools may not detect. This coincides with Agile Software Development guidelines (Hinderks et al., 2022) where quick evaluation allows to inform the next iteration of the development process.

Our proof of concept study (Section 3) illustrates the need to develop easy-to-use emotion trackers that can estimate emotions related to short-timed events. Such trackers would improve the sensitivity of emotional state measuring devices and would make it possible to monitor emotions in numerous research settings where it is currently unfeasible.

As a final thought, we believe that much research is still needed to unite our current understanding of ethnomusicology, music theory, and human musical practice with the technological advances being made in generative AI, to create an ethical, inclusive, diverse technology that will pave the way for the making of music in the future.

6. Limitations of the study

This article presents a comprehensive methodology for user-based evaluation of AI-generated music. The methodology proposed uses human evaluation, which can be tailored specifically to the object of study, and provides several alternatives for elaborating surveys, gathering emotional feedback from participants, using different kinds of generators and musical pieces, and analyzing data. Although most state-of-the-art generators use non-standardized versions of some of the evaluation tools proposed and could even fit into the proposed methodology with an adequate selection of options in the Adaptability Decision Tree and Matrices, a systematic review of all possible resources for AI music evaluation is beyond the scope of this study.

One of the main objective of this study is to provide future researchers with the tools to formulate AI music generation evaluation studies tailored to their specific contexts and needs, while still providing common ground for standardization and comparison between studies. Nevertheless, researchers using this methodology should be aware of the potential biases towards western music these methodology may have.

Moreover, the provided strategies and tools to assess emotion in music following guidelines that suggest the need for emotion validation using multiple approaches (Eerola & Vuoskoski, 2013) (through self-assessment, emotion detection

and physiological measurements). These approaches can coincide with both continuous and discrete emotion models. They are very useful when mixed with objective validation methods (Subsection 1.1) to improve the reliability of datasets labeling and model refinement, as these approaches are easily translated into numerical formulas. On the other hand, researchers are advised against taking these emotional estimations as absolute measurements. Music emotions depends on large myriad of factors that are very subjective and are not completely covered by the basic emotion or the affective circumplex models (Eerola & Saari, 2025). Context, where the music is made and received outside the laboratories, is also to be accounted for.

The article also provides a proof-of-concept in which the methodology is applied and in which many of the proposed tools are used. This proof of concept throws light on the applicability of the tools and can aid future researchers in designing a benchmark for their evaluation studies. Due to the small number of participants, however, the potentially interesting implications of its results should be further studied. This pilot study shows how the methodology can be applied for AI generation. It should be noted that AI generation is a very rapidly evolving field, and that the generators used in the study may become obsolete in the near future. Nevertheless, the necessity for constant evaluation and supervision remain; while the perspective gained through the study can point towards a much wider acceptance for the technology, with the use of next-generation systems in the near future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethics statement

This study was reviewed and approved by the ethical Committee of Universidad Loyola.

Funding

This research was supported by the EQUAVEL project PID2022-137646OB-C31, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.

ORCID

Miguel Civit  <http://orcid.org/0000-0003-4310-6377>
Véronique Drai-Zerbib  <http://orcid.org/0000-0002-5623-6229>
Francisco Cuadrado  <http://orcid.org/0000-0003-2307-3846>
Maria J. Escalona  <http://orcid.org/0000-0002-6435-1497>

Data availability statement

Music Excerpts from the pilot study are available in their respective datasets and in: *Music Folder*

References

Agawu, K. (2006). Structural analysis or cultural analysis? competing perspectives on the “standard pattern” of west african rhythm.

- Journal of the American Musicological Society*, 59(1), 1–46. <https://doi.org/10.1525/jams.2006.59.1.1>
- Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.
- Ahmad, Z., & Khan, N. (2022). A survey on physiological signal-based emotion recognition. *Bioengineering*, 9(11), 688. <https://doi.org/10.3390/bioengineering9110688>
- Ariza, C. (2009). The interrogator as critic: The turing test and the evaluation of generative music systems. *Computer Music Journal*, 33(2), 48–70. <https://doi.org/10.1162/comj.2009.33.2.48>
- Bittner, R. M., Bosch, J. J., Rubinstein, D., Meseguer-Brocal, G., & Ewert, S. (2022). A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- Briot, J.-P., & Pachet, F. (2020). Deep learning for music generation: Challenges and directions. *Neural Computing and Applications*, 32(4), 981–993. <https://doi.org/10.1007/s00521-018-3813-6>
- Bugnon, L. A., Calvo, R. A., & Milone, D. H. (2020). Dimensional affect recognition from HRV: An approach based on supervised SOM and ELM. *IEEE Transactions on Affective Computing*, 11(1), 32–44. <https://doi.org/10.1109/TAFFC.2017.2763943>
- Cayrou, S., Dicks, P., Gauvain-Piquard, A., Dolbeault, S., Callahan, S., & Roge, B. (2000). Validation de la traduction française du poms (profile of mood states). *Psychologie et psychométrie*, 21(4), 5–22.
- Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.
- Chu, H., Kim, J., Kim, S., Lim, H., Lee, H., Jin, S., & Ko, S. (2022). An empirical study on how people perceive ai-generated music. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 304–314).
- Chuan, C.-H., Agres, K., & Herremans, D. (2020). From context to concept: Exploring semantic relationships in music with word2vec. *Neural Computing and Applications*, 32(4), 1023–1036. <https://doi.org/10.1007/s00521-018-3923-1>
- Cideron, G., Girgin, S., Verzetti, M., Vincent, D., Kastelic, M., & Borsos, Z. (2024). Musicrl: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229*.
- Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J. (2022). A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*, 209, 118190. <https://doi.org/10.1016/j.eswa.2022.118190>
- Civit, M., Drai-Zerbib, V., Lizcano, D., & Escalona, M. J. (2024). Sunocaps: A novel dataset of text-prompt based ai-generated music with emotion annotations. *Data in Brief*, 55, 110743. <https://doi.org/10.1016/j.dib.2024.110743>
- Civit, M., Escalona, M. J., Cuadrado, F., & Reyes-de Cozar, S. (2024). Class integration of chatgpt and learning analytics for higher education. *Expert Systems*, 41(12), e13703. <https://doi.org/10.1111/exsy.13703>
- Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., & Défossez, A. (2024). Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 47704–47720. <https://doi.org/10.48550/arXiv.2306.05284>
- Cuadrado, F., Lopez-Cobo, I., Mateos-Blanco, T., & Tajadura-Jiménez, A. (2020). Arousing the sound: A field study on the emotional impact on children of arousing sound design and 3d audio spatialization in an audio story. *Frontiers in Psychology*, 11, 737. <https://doi.org/10.3389/fpsyg.2020.00737>
- Cuculo, V., & D’Amelio, A. (2019). Openfacs: An open source facs-based 3d face animation system. In *Image and graphics: 10th international conference, ICIG 2019, Beijing, China, August 23–25, 2019, proceedings, part ii 10* (pp. 232–242).
- Dervakos, E., Filandrianos, G., & Stamou, G. (2021). Heuristics for evaluation of ai generated music. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 9164–9171).
- Dimitriev, D., Indeykina, O., & Dimitriev, A. (2022). The effect of auditory stimulation on the nonlinear dynamics of heart rate: The impact of emotional valence and arousal. *bioRxiv*, 2022–2003.

- Diwanji, V. S., Geana, M., Pei, J., Nguyen, N., Izhar, N., & Chaif, R. H. (2025). Consumers' emotional responses to ai-generated versus human-generated content: The role of perceived agency, affect and gaze in health marketing. *International Journal of Human-Computer Interaction*, 1–21. <https://doi.org/10.1080/10447318.2025.2454954>
- Dolan, J. G. (2008). Shared decision-making-transferring research into practice: The analytic hierarchy process (AHP). *Patient Education and Counseling*, 73(3), 418–425. <https://doi.org/10.1016/j.pec.2008.07.032>
- Dong, H.-W., Chen, K., McAuley, J., & Berg-Kirkpatrick, T. (2020). Muspy: A toolkit for symbolic music generation. In *Proceedings of the 21st International Society for Music Information Retrieval conference (ISMIR)*.
- Eerola, T., & Saari, P. (2025). What emotions does music express? structure of affect terms in music using iterative crowdsourcing paradigm. *PLoS One*, 20(1), e0313502. <https://doi.org/10.1371/journal.pone.0313502>
- Eerola, T., & Vuoskoski, J. K. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception*, 30(3), 307–340. <https://doi.org/10.1525/mp.2012.30.3.307>
- Eerola, T., Vuoskoski, J. K., Peltola, H.-R., Putkinen, V., & Schäfer, K. (2018). An integrative review of the enjoyment of sadness associated with music. *Physics of Life Reviews*, 25, 100–121. <https://doi.org/10.1016/j.plrev.2017.11.016>
- Ferreira, P., Limongi, R., & Fávero, L. P. (2023). Generating music with data: Application of deep learning models for symbolic music composition. *Applied Sciences*, 13(7), 4543. <https://doi.org/10.3390/app13074543>
- Frid, E., Gomes, C., & Jin, Z. (2020). Music creation by example. In *Proceedings of the 2020 Chi conference on human factors in computing systems* (pp. 1–13).
- García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What do we mean by genai? A systematic mapping of the evolution, trends, and techniques involved in generative ai. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(4), 7. <https://doi.org/10.9781/ijimai.2023.07.006>
- Goepel, K. D. (2018). Implementation of an online software tool for the analytic hierarchy process (ahp-os). *International Journal of the Analytic Hierarchy Process*, 10(3), 590. <https://doi.org/10.13033/ijahp.v10i3.590>
- Gutknecht, M., Danner, M., Schaarschmidt, M.-L., Gross, C., & Augustin, M. (2018). Assessing the importance of treatment goals in patients with psoriasis: Analytic hierarchy process vs. likert scales. *The Patient*, 11(4), 425–437. <https://doi.org/10.1007/s40271-018-0300-1>
- Hadjeres, G., Pachet, F., & Nielsen, F. (2017). Deepbach: A steerable model for bach chorales generation. In *International conference on machine learning* (pp. 1362–1371).
- Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., & Eck, D. (2019). Enabling factorized piano music modeling and generation with the maestro dataset. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*.
- Hernández-Orallo, J. (2020). Twenty years beyond the turing test: Moving beyond the human judges too. *Minds and Machines*, 30(4), 533–562. <https://doi.org/10.1007/s11023-020-09549-0>
- Hernando, D., Roca, S., Sancho, J., Alesanco, Á., & Bailón, R. (2018). Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects. *Sensors*, 18(8), 2619. <https://doi.org/10.3390/s18082619>
- Hinderks, A., Mayo, F. J. D., Thomaschewski, J., & Escalona, M. J. (2022). Approaches to manage the user experience process in agile software development: A systematic literature review. *Information and Software Technology*, 150, 106957. <https://doi.org/10.1016/j.infsof.2022.106957>
- Hirten, R. P., Danieleto, M., Tomalin, L., Choi, K. H., Zweig, M., Golden, E., Kaur, S., Helmus, D., Biello, A., Pyzik, R., Calcagno, C., Freeman, R., Sands, B. E., Charney, D., Bottinger, E. P., Murrrough, J. W., Keefer, L., Suarez-Farinas, M., Nadkarni, G. N., & Fayad, Z. A. (2021). Factors associated with longitudinal psychological and physiological stress in health care workers during the covid-19 pandemic: Observational study using apple watch data. *Journal of Medical Internet Research*, 23(9), e31295. <https://doi.org/10.2196/31295>
- Huang, J., Wang, J.-C., Smith, J. B., Song, X., & Wang, Y. (2021). Modeling the compatibility of stem tracks to generate music mash-ups. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 187–195).
- Ing, E. B. (2021). A survey-weighted analytic hierarchy process to quantify authorship. *Advances in Medical Education and Practice*, 12, 1021–1031. <https://doi.org/10.2147/AMEP.S328648>
- Ismail, S. N. M. S., Aziz, N. A. A., Ibrahim, S. Z., & Mohamad, M. S. (2024). A systematic review of emotion recognition using cardio-based signals. *ICT Express*, 10(1), 156–183.
- Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N., & Miyakawa, R. H. (2020). Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 516–520).
- Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3), 246–279. <https://doi.org/10.1007/s12559-012-9156-1>
- Katz, B. (2015). Sound board: Can we stop the loudness war in streaming? *Journal of the Audio Engineering Society*, 63(11), 939–940. <https://aes2.org/publications/elibrary-page/?id=18053>
- Katz, B., & Katz, R. A. (2003). *Mastering audio: The art and the science*. Butterworth-Heinemann.
- Kayser, D., Egermann, H., & Barraclough, N. E. (2022). Audience facial expressions detected by automated face analysis software reflect emotions in music. *Behavior Research Methods*, 54(3), 1493–1507. <https://doi.org/10.3758/s13428-021-01678-3>
- Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 2350–2354). <https://doi.org/10.21437/Interspeech.2019-2219>
- Knejzliková, T., Světlák, M., Malatincová, T., Roman, R., Chládek, J., Najmanová, J., Theiner, P., Linhartová, P., & Kašpárek, T. (2021). Electrodermal response to mirror exposure in relation to subjective emotional responses, emotional competences and affectivity in adolescent girls with restrictive anorexia and healthy controls. *Frontiers in Psychology*, 12, 673597. <https://doi.org/10.3389/fpsyg.2021.673597>
- Kodra, E., Senechal, T., McDuff, D., & El Kaliouby, R. (2013). From dials to facial coding: Automated detection of spontaneous facial expressions for media research. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1–6).
- Kulke, L., Feyerabend, D., & Schacht, A. (2020). A comparison of the affectiva imotions facial expression analysis software with emg for identifying facial expressions of emotion. *Frontiers in Psychology*, 11, 329. <https://doi.org/10.3389/fpsyg.2020.00329>
- Lui, G. Y., Loughane, D., Polley, C., Jayarathna, T., & Breen, P. P. (2022). The apple watch for monitoring mental health-related physiological symptoms: Literature review. *JMIR Mental Health*, 9(9), e37354. <https://doi.org/10.2196/37354>
- McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016). Validation of the empathica e4 wristband. In *2016 IEEE Embs International Student Conference (ISC)* (pp. 1–4).
- Moysis, L., Iliadis, L. A., Sotiroidis, S. P., Kokkinidis, K., Sarigiannidis, P., & Nikolaidis, S. (2023). The challenges of music deep learning for traditional music. In *2023 12th international conference on modern circuits and systems technologies (Mocast)* (pp. 1–5).
- Muñoz-Saavedra, L., Escobar-Linero, E., Miró-Amarante, L., Bohórquez, R., & M, D.-M. (2023). Designing and evaluating a wearable device for affective state level classification using machine learning techniques. *Expert Systems with Applications*, 219, 119577. <https://doi.org/10.1016/j.eswa.2023.119577>
- Osoba, O. A., Welser, W. IV., & Welser, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.
- Ponsiglione, A. M., Amato, F., Cozzolino, S., Russo, G., Romano, M., & Improta, G. (2022). A hybrid analytic hierarchy process and likert scale

- approach for the quality assessment of medical education programs. *Mathematics*, 10(9), 1426. <https://doi.org/10.3390/math10091426>
- Ribeiro, F., Florêncio, D., Zhang, C., & Seltzer, M. (2011). Crowdmoss: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2416–2419).
- Rosenberg, E. L., & Ekman, P. (2020). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press.
- Saaty, T. L., & Özdemir, M.S. (2014). How many judges should there be in a group? *Annals of Data Science*, 1, 359–368. <https://doi.org/10.1007/s40745-014-0026-4>
- Schippers, A., Aben, B., Griep, Y., & Van Overwalle, F. (2018). Ultra-short term heart rate variability as a tool to assess changes in valence. *Psychiatry Research*, 270, 517–522. <https://doi.org/10.1016/j.psychres.2018.10.005>
- Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception*, 21(4), 561–585. <https://doi.org/10.1525/mp.2004.21.4.561>
- Skiendziel, T., Rösch, A. G., & Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring. *PLoS One*, 14(10), e0223905. <https://doi.org/10.1371/journal.pone.0223905>
- Song, B., & Kang, S. (2016). A method of assigning weights using a ranking and nonhierarchy comparison. *Advances in Decision Sciences*, 2016, 1–9. <https://doi.org/10.1155/2016/8963214>
- Sturm, B. L., & Ben-Tal, O. (2017). Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal of Creative Music Systems*, 2(1), 32–60. <https://doi.org/10.5920/JCMS.2017.09>
- Sullivan, G. M., & Artino, A. R. Jr. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, 5(4), 541–542. <https://doi.org/10.4300/JGME-5-4-18>
- Tanoue, Y., Nakashima, S., Komatsu, T., Kosugi, M., Kawakami, S., Kawakami, S., Michishita, R., Higaki, Y., & Uehara, Y. (2023). The validity of ultra-short-term heart rate variability during cycling exercise. *Sensors*, 23(6), 3325. <https://doi.org/10.3390/s23063325>
- Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios hrv–heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, 113(1), 210–220. <https://doi.org/10.1016/j.cmpb.2013.07.024>
- Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. In *2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1934–1937).
- Wang, S., Bao, Z., & Jingtong, E. (2021). Armor: A benchmark for meta-evaluation of artificial music. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 5583–5590).
- Wiggins, G. A. (2019). A framework for description, analysis and comparison of creative systems. In T. Veale & F. Cardoso (Eds.), *Computational creativity*. Computational synthesis and creative systems. Springer. https://doi.org/10.1007/978-3-319-43610-4_2
- Xiong, Z., Wang, W., Yu, J., Lin, Y., & Wang, Z. (2023). A comprehensive survey for evaluation methodologies of ai-generated music. arXiv preprint arXiv:2308.13736.
- Yang, L.-C., & Lerch, A. (2020). On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9), 4773–4784. <https://doi.org/10.1007/s00521-018-3849-7>

About the authors

Miguel Civit has a PHD in Computer Science from the Université de Bourgogne and University of Seville with a focus on quantitative and qualitative validation of AI in Music. His research focuses on the effects of emotion in AI music generation. He is a member of the ES3 research group.

Véronique Drai-Zerbib is a professor of cognitive psychology at the University of Bourgogne Europe, specializing in expertise development, musical cognition, expert memory, multimodal information integration, digital reading, musical reading. She uses behavioral and neurophysiological approaches such as eye-tracking and also virtual reality and machine learning.

Francisco Cuadrado is PhD in Communication, researcher, professor, composer and sound designer. His research fields are music and sound creation and perception in media, and social and emotional development through music. He has been the IP for different research projects, like “Learning To Be” and “The Unconscious Listening.”

María José Escalona, PhD in Computer Engineering from the University of Seville, is a full professor and director of the ES3 research group. Specializing in web engineering and software quality, she developed the NDT methodology, widely used in industry. She has an extensive research career, with numerous publications and projects.