
Rotterdam School of Management

MSc Business Information Management

Master's Thesis

---

# Do We Still Need Taylor Swift?

## An Experiment Testing Evaluations of Music Featuring AI-Cloned Vocals

---

Gabriele Landi

657675

Thesis Coach: Agnieszka Kloc

Co-reader: Dr. Dominik Gutt

July 16th, 2024

## Preface

## Acknowledgments

First and foremost, I would like to express my sincere gratitude to my thesis coach, Agnieszka Kloc. We started this journey together two years ago with little idea of the path it would take. While your invaluable feedback was crucial to the completion of this work, I am most thankful for your patience, understanding, and support through the ups and downs of this journey. I couldn't have asked for a better coach.

I also wish to thank my co-reader and former boss, Dr. Dominik Gutt. Although our discussions about the thesis were limited, your insights were vital to its validity. Additionally, your course on research methods was instrumental in shaping the methodology.

I've often heard that during your studies, you form friendships that last a lifetime. It always sounded like a comforting, almost idealised narrative to lighten the weight of academic life, and I never quite believed it. As it turns out, the past two years have shown me just how wrong I was. I've crossed paths with people who have made, and continue to make, a meaningful difference in my life.

A special appreciation goes to Szymon. While a random text message brought us together, your friendship has been anything but random. So much of who I am today is because of you. Keep it sexy, keep it slay. Another heartfelt thanks goes to Zuza. Your friendship and our gossip-filled chats have been priceless. Through it all, you two were always there - whether it was partying and sharing the fun times, or offering me shelter when I needed it most.

I also owe a great deal to Fanny, whose supply of Red Bull has been key to my writing, and to Piotr, a friend first and flatmate second. Many thanks as well to all the Sexi Darlings out there.

I'm also grateful to everyone who has been by my side during the past two years. A special mention goes to Amita and Edoardo, friends for as long as I can remember, and to Chiara, for a rekindled friendship.

A note of appreciation goes to Zia Ale, for always remembering my name-day and being the ultimate 'behind-the-scenes' support I can rely on.

My deepest appreciation goes to my family. This achievement is as much theirs as it is mine. Their unwavering support has been my greatest source of strength and motivation.

To Carolina, cousin in name but sister at heart, thank you for being a constant source of laughter and inspiration. A bright (and wealthy) future awaits us both! To Zia Stefania, who has been the extra dash of spice in these past few years.

To my parents, I am forever grateful for the example you both have set, and I can only hope to live up to the values you've instilled in me. To my mother, whose strength and resilience inspire me every day: thank you for your endless love, for always putting us before yourself, for believing in me and for teaching me to aim high and never settle for anything less than I deserve. To my father, the most selfless person I know: thank you for making this journey possible. Your willingness to give and sacrifice so much for our family is a constant reminder of what true dedication and love look like.

To my triplet siblings, Riccardo and Ludovica, and my sister, Caterina, thank you for the love you have shown and continue to show throughout these years; I might not always express it, but I am forever grateful to have you in my life, shaping the person I am today. I look up to you all. To Riccardo, as brilliant as one can be, who, perhaps unknowingly, has unlocked so much of who I am today, what I believe in, and who I aspire to be. Thank you for showing me that sometimes, stepping back is the key to moving forward. To Ludovica, thank you for adding that touch of wildness and chaos I always needed. All we've shared together, the highs and lows, has been key to everything I've achieved. Without you, I would have been lost more times than I can count. To Caterina, thank you for being the 'fourth' triplet and the glue that holds us all together. Words fall short when it comes to describing your big heart and boundless patience. Also, thank you for guiding me through the 'wonderful' world of LaTeX.

Finché l'anima balla

## Executive Summary

In an era where Artificial Intelligence (AI) shapes how music is produced, shared, and experienced, understanding public reactions to its applications is of paramount importance. AI-generated vocals have emerged in contentious scenarios, such as Drake's recent diss track using Tupac's cloned vocals, raising questions about the public's readiness to embrace this technology. By drawing parallels with AI instrumental music and leveraging insights from research within the creative domain, this paper examines the elements influencing public evaluations of music featuring AI-cloned vocals.

People's attitudes and expectations are decisive in determining whether AI-cloned vocals will set a new trend in music production and consumption. Indeed, while artistic merit has the potential to transcend the label 'AI created', attitudes and expectations are pivotal in the acceptance and appreciation of musical compositions. Based on this premise, our study is centred around the following research question: how do expectations about songs featuring AI-cloned vocals influence people's evaluation of such songs?

As with any emerging technology, people are destined to be positively or negatively surprised by how AI-cloned vocals compare to their beliefs. Understanding how this translates to the evaluation of music is crucial in determining how AI-generated vocals can be integrated into the industry. Originally centred in the domain of interpersonal communication, Expectancy Violation Theory (EVT) has evolved to include human-computer interactions (HCI), presenting a relevant framework to address how expectations influence the evaluation of songs with AI-cloned vocals. This research contributes to EVT by offering insights into its application to AI singers and determining whether it extends similarly to AI musicians. Accordingly, our study investigates the following research question: how does violating expectations about songs with AI-cloned vocals influence people's evaluation of such songs?

Adopting an experimental approach, this study incorporates a survey experiment framed within a 2x2x2 factorial design. Specifically, the study manipulates expectations (positive vs. negative) and vocal quality (high vs. low) across two music genres (pop and rock) to uncover cause-and-effect relationships between these factors and the evaluation of AI vocal music. The

findings indicate that vocal quality significantly outweighs other elements when evaluating pop and rock music. Indeed, the evaluations of AI vocal music remain unaffected by initial expectations and whether the vocal quality was much better or worse than anticipated. Instead, people evaluate songs positively if the AI-cloned vocals are of high quality and negatively if they are of low quality, regardless of expectations, their violations, and music genre. This diverges from the patterns observed in AI instrumental music, suggesting that the introduction of AI vocals set in motion novel dynamics in how listeners evaluate AI music. Moreover, the study uncovers a general negative sentiment towards AI singers, suggesting that people may view singing as a uniquely human trait that machines should not replicate.

The study contributes to academia by addressing the literature gap in public perceptions of AI applications within the creative domain, adding to the discussion of human-computer interaction within the music domain, and offering insights into how EVT applies to AI vocal music. Additionally, the research puts forth a new proposition that varying levels of acceptance for AI musicians and AI singers might be linked to the human-like characteristics inherent to each role. Simultaneously, it offers important managerial implications, aiding music producers, technology developers, and marketers navigate the initial stages of integrating AI-cloned vocals into the industry.

Nonetheless, it is important to acknowledge that this study has certain limitations. The technology has only recently advanced to enter the public domain. While experimental research is particularly useful for testing theories, future studies might adopt a more exploratory approach using qualitative methods to establish foundational elements central to the domain of AI-cloned vocals. The experimental design also has inherent limitations; for instance, evaluating music in a controlled setting may not fully represent real-world listening experiences. Additionally, the study targets a specific population, which limits the generalisability of these findings to a particular group within the European Union. Despite these limitations, this research acts as a stepping stone for future studies, providing the first empirical evidence of public reactions to music featuring AI-cloned vocals.

## Table of Contents


**1 Introduction** **13**

1.1 Relevance . . . . . 14

1.2 Research Gap . . . . . 17

1.3 Research Questions . . . . . 19

1.4 Research Design Overview . . . . . 19

1.5 Research Outline . . . . . 20

**2 Literature Review** **20**

2.1 Voice Cloning: Origins and Evolution . . . . . 21

2.1.1 The Mechanical, Electrical, and Digital Era . . . . . 21

2.1.2 Voice Cloning: Text-to-Speech . . . . . 23

2.1.3 Voice Cloning: Speech-to-Speech . . . . . 25

2.1.4 Real-World Applications . . . . . 26

2.1.5 Applications in the Music Industry . . . . . 27

2.2 Public Perceptions of Creative AI . . . . . 29

2.2.1 Hypotheses Development . . . . . 30

2.3 Expectancy Violation Theory . . . . . 31

2.3.1 Hypotheses Development . . . . . 33

2.4 Music Genre . . . . . 34

2.5 Conceptual Model . . . . . 37

**3 Methodology** **37**

3.1 Research Design . . . . . 37

3.1.1 Experimental Design . . . . . 38

3.2 Data Collection . . . . . 39

3.3 Experimental Procedure . . . . . 40

3.3.1 Pre-Survey Briefing . . . . . 40

3.3.2 Baseline Measures . . . . . 40

3.3.3 Expectations Manipulation . . . . . 41

3.3.4 Manipulation Check: Expectations . . . 41
3.3.5 Vocals Quality & Music Genre Manipulations . . . 42
3.3.6 Manipulation Check: Vocals Quality . . . 43
3.3.7 Demographics & Remarks . . . 44
3.3.8 Attention Checks . . . 44
3.3.9 Survey Design Elements to Mitigate Bias . . . 44
3.4 Sample Size . . . 45
3.5 Sample Population . . . 46
3.6 Operationalisation of Constructs . . . 47
3.6.1 Dependent Variable . . . 47
3.6.2 Independent Variables . . . 48
3.6.3 Moderating Variables . . . 49
3.6.4 Control Variables & Music Genre . . . 49
3.7 Data Analysis . . . 51

**4 Preliminary Analysis** **53**
4.1 Sample Demographics . . . 53
4.2 Scales Reliability . . . 54
4.3 Scales Validity . . . 55
4.4 Randomisation Check . . . 57
4.5 Manipulation Checks . . . 58
4.6 Visual Inspection . . . 61
4.7 Descriptive Statistics . . . 62

**5 Results** **65**
5.1 Preliminary Testing with Simple Linear Regressions . . . 65
5.1.1 Assumptions: Simple Linear Regression . . . 66
5.1.2 Results: Simple Regression Analysis . . . 67
5.1.3 Robustness Check: Bootstrapping . . . 69
5.1.4 Robustness Check: Spearman's Rank Correlation . . . 69

5.2 Hypotheses Testing . . . 69
5.2.1 Assumptions: Multiple Linear Regression . . . 70
5.2.2 Results: Multiple Regression Analysis . . . 71
5.2.3 Robustness Check: ANCOVA . . . 75
5.2.3.1 Assumptions: ANCOVA . . . 75
5.2.3.2 Results: ANCOVA . . . 76
5.2.4 Robustness Check: Hayes' PROCESS Macro . . . 78

**6 Discussion** **79**
6.1 Findings and Implications . . . 79
6.1.1 Academic Implications . . . 82
6.1.2 Managerial Implications . . . 83

**7 Limitations & Future Research** **84**
7.1 Validity & Generalisability . . . 84
7.2 Econometric Models . . . 87
7.3 Future Directions . . . 87

**8 References** **89**

**9 Appendix** **103**
9.1 Appendix A . . . 103
9.2 Appendix B . . . 105
9.3 Appendix C . . . 105
9.4 Appendix D . . . 108
9.5 Appendix E . . . 110
9.6 Appendix F . . . 110
9.7 Appendix G . . . 112
9.8 Appendix H . . . 115
9.9 Appendix I . . . 116
9.10 Appendix J . . . 118

9.11 Appendix K . . . . . 119
9.12 Appendix L . . . . . 122
9.13 Appendix M . . . . . 123
9.14 Appendix N . . . . . 125
9.15 Appendix O . . . . . 126
9.16 Appendix P . . . . . 127
9.17 Appendix Q . . . . . 129
9.18 Appendix R . . . . . 130
9.19 Appendix S . . . . . 131
9.20 Appendix T . . . . . 133

## Table of Tables


1 List of 14 Music Genres . . . . . 35
2 Scales Reliability . . . . . 55
3 Exploratory Factor Analysis (Pop) . . . . . 56
4 Random assignment Check - Mann-Whitney U Test . . . . . 58
5 Random assignment Check - Chi-Squared Test . . . . . 58
6 Manipulation Check - Expectations: Article 1 vs. Article 2 . . . . . 59
7 Manipulation Check - Expectations: Article 1 vs. Control . . . . . 59
8 Manipulation Check - Expectations: Article 2 vs. Control . . . . . 59
9 Manipulation Check - Pop - High vs. Low Quality . . . . . 60
10 Manipulation Check - Rock - High vs. Low Quality . . . . . 60
11 Manipulation Check - Expectancy violation high-quality samples . . . . . 61
12 Manipulation Check - Expectancy violation low-quality samples . . . . . 61
13 Song Samples Evaluation . . . . . 62
14 Descriptive Statistics Expectancy Violation . . . . . 63
15 Descriptive Statistics Control & Covariates . . . . . 64
16 Simple Linear Regression Models . . . . . 66
17 Simple Linear Regression Models Output . . . . . 68
18 Multiple Regression Models . . . . . 70
19 Multiple Linear Regression Models Output . . . . . 74
20 ANCOVA Output Pop . . . . . 77
21 ANCOVA Output Rock . . . . . 77
22 Hypotheses Overview . . . . . 80

## Table of Figures


1 The typical pipeline of the latest TTS systems . . . . . 24

2 The typical pipeline of a voice conversion system . . . . . 26

3 Conceptual model . . . . . 37

4 Overview of the 2x2x2 factorial design . . . . . 38

5 Overview of the song samples matrix . . . . . 43

6 Survey experimental flow . . . . . 52

7 ANCOVA interaction plots . . . . . 76

# 1 Introduction

AI stands at the centre of an ongoing debate, with people holding contrasting views on recent trends and their implications (Hong and Curran, 2019). Advocates expect AI to revolutionise problem-solving and information processing, ultimately enhancing efficiency, decision-making, and cost-effectiveness (Makridakis, 2017). In contrast, sceptics express concerns about the potential for AI systems to escalate beyond control, produce biased or unethical results, and cause job displacement (Fast and Horvitz, 2017). While AI taking over tasks traditionally performed by humans - to optimize processes and reduce costs – is already prevalent across multiple industries (Huang and Rust, 2018), it is becoming increasingly relevant in the creative domains within the art and music fields (Anantrasiricha and Bull, 2021).

In the art field, AI technologies play a role in analysing existing art, detecting forged artworks, and facilitating the creation and co-creation of new art pieces (Cetinic and She, 2022; Latika et al., 2023). In the music industry, AI is transforming how music is produced, shared, and experienced. For instance, streaming platforms use AI algorithms to recommend music based on listening habits (Shank et al., 2023). Companies such as LANDR provide AI-driven mixing and mastering technologies, leading to the democratisation of music production by making these processes cheaper, faster, and more accessible (Birtchnell and Elliott, 2018). AI tools also assist in the generation of lyrics, the co-production of music, and the composition of complete musical works, opening up new avenues for artistic expression and challenging traditional definitions of music (Shank et al., 2023). Finally, voice cloning, a relatively new AI technology, is poised to become a disruptive force in the music industry. This technology uses artificial intelligence to replicate a person's voice, allowing for the creation of natural-sounding audio samples. It can capture the pitch and tone as well as the emotional nuances and distinct traits of the reference voice (Weitzman, 2023).

The latest and most striking example of voice cloning is 'Heart on My Sleeve', an AI-generated song that clones the vocals of the famous artists Drake and The Weeknd without their consent (Coscarelli, 2023). The song went viral on social media and streaming platforms in the early months of 2023, gaining millions of streams and views before being taken down

(Snapes, 2023). Among other popular artists who have fallen victim to viral unauthorised voice cloning are AI Rihanna singing 'Cuff It' by Beyoncé and AI Kanye West performing 'Hey There Delilah'. In light of these cases, AI voice cloning has earned immediate comparisons to earlier disruptive technologies in the music industry, such as synthesisers (Coscarelli, 2023). The technology stands to redefine the social role of singers, potentially revive the voices of past artists, and democratise music production by making it easier, faster, and more cost-effective. For instance, AI voice cloning can replace costly voice actors and singers for background vocals (Weitzman, 2023). However, it also raises intricate legal and ethical concerns. For instance, a voice cannot be copyrighted (Harkins, 2012), yet it is unlawful to attribute a song to someone without their explicit consent. Additionally, using a deceased artist's voice inevitably elicits debates surrounding consent and legacy (Kim, 2023).

To date, industry stakeholders have taken different positions. Some maintain an optimistic stance, arguing that it is not the end of art but rather marks the birth of a new musical genre that could transform the industry, provided that protections and legal frameworks are updated (Kim, 2023). A few early adopters, such as musician Holly Herndon, have already consented to the production of songs using their vocal samples, with a system for compensation in place (Coscarelli, 2023). On the other hand, entities like Universal Music Group, the world's leading music company (Shevlin and Shan, 2019), perceive it 'as a threat to the human creative expression'. They argue it primarily fuels deep fakes and fraud while preventing artists from receiving their rightful compensation (Ibrahim, 2023). At this stage, it is difficult to make predictions about the future of the technology, as many elements still need to be addressed before either viewpoint can prevail. This research aims to address some of these elements by exploring the role expectations and vocal quality play in shaping public perception and acceptance of this emerging technology.

## 1.1 Relevance

While it is increasingly expected for AI to outperform humans in computational tasks (Hong et al., 2021), its applications in the creative sector, including the music industry, and its involvement in creative and artistic endeavours inherently challenge contemporary notions of

creativity (Hong and Curran, 2019). However, Coeckelbergh (2017) argues that if art is defined using objective criteria, such as harmony or rhythm, AI could be engineered to produce works that meet these criteria. Alternatively, if the definition of art relies on subjective criteria, then any creation, including those generated by AI, might qualify as art. This distinction shifts the focus from questioning whether songs that clone the vocals of artists fit within the traditional scope of art to considering if such songs can be valued and accepted on par with human-created songs and what factors contribute to their evaluation or intention to listen to them.

In this regard, different attitudes towards creative AI results in a different evaluation of its products. While an open-minded attitude leads to a positive evaluation, unwillingness to embrace AI products may hinder their appreciation (Hong et al., 2021). In the field of artificial intelligence and music, the label 'AI created' may prove less significant in the face of artistic merit, with research indicating that listeners' appreciation of a musical piece is primarily influenced by its quality rather than its origin (Hong et al., 2022). However, listeners may harbour biases against music that they think was created by AI if it falls short of their expectations of what AI is capable of creating (Shank et al., 2023). Research also indicates that expectations and their violations shape people's attitudes toward AI products, and thus influence their valuation of AI music (Hong et al., 2022). Essentially, while artistic merit has the potential to transcend the label 'AI created', attitudes and expectations play a crucial role in the reception and appreciation of musical compositions. Therefore, while this study acknowledges the numerous aspects that could be addressed, it specifically focuses on the role of expectations and their impact on the appreciation of music featuring AI-generated vocals, as these factors are pivotal in understanding public acceptance of this emerging technology.

Typically, people form expectations about their experience using AI products and evaluate them based on whether these expectations are met or violated (Burgoon et al., 2016). However, as AI technology rapidly evolves, what is known and expected of AI today may soon be outdated. Additionally, people may form expectations based on media content, but how media presents the technology only sometimes aligns with reality (Aïmeur et al., 2023). Therefore, regardless of people's expectations about AI products, at some point, new developments will either exceed or fall short of these expectations, influencing their evaluation and intention

to adopt them (Hong, 2021).

Voice cloning technology is no exception. The technology has recently undergone significant advancements, leading to improvements in the quality of generated voices (Masood et al., 2023). Its applications in the music industry are just beginning to surface, leaving public expectations and reactions to its use in songs somewhat unclear and unexplored. As this technology becomes more widespread, expectations are also likely to change. Thus, while industry stakeholders are keen to measure the public reception of these songs and their intention to listen to them, they first need to understand how expectations influence such outcomes. Expectancy Violation Theory (EVT) offers a relevant framework to understand and predict how people respond when their expectations are met or violated (Burgoon et al., 2016). This research aims to apply EVT in the context of music featuring AI-cloned vocals, thus making a significant contribution to developers, marketers, and other industry stakeholders interested in anticipating and managing individuals' expectations more effectively.

The perception and value placed on voice cloning technology may also vary across different contexts. Specifically, in high-involvement contexts where listeners are deeply focused on the music, it might be more challenging to introduce AI-cloned vocals compared to situations of low listener involvement, where the music plays a more background role (Moura and Maw, 2021). This trend is already observable for AI-composed music, with a growing acceptance of its use in low-involvement contexts such as commercials, public spaces, or political campaigns (Moura and Maw, 2021). Researchers and industry stakeholders should investigate how listeners perceive voice cloning technology in different contexts before launching it commercially. By measuring individuals' evaluation of music in a controlled setting, this research seeks to provide insights to marketers and industry stakeholders on how AI-cloned vocals are perceived in high-involvement contexts.

In conclusion, this research's managerial relevance highlights the importance for technology developers and industry stakeholders to understand expectations about AI-cloned vocals. This understanding will not only aid in the development of appropriate regulations but also unveil the true economic potential of songs featuring AI-cloned vocals. By better anticipating market trends, optimising marketing strategies, and ensuring that the quality of AI-cloned vo-

cals aligns with public acceptance and appreciation, stakeholders may ultimately contribute to the successful integration of AI-cloned vocals into the music industry.

## 1.2 Research Gap

In recent years, applications of AI within the creative domain have grown significantly (Anantrasirichai and Bull, 2021), with research publications increasing by more than 500% in several countries (Davies, 2020). However, our understanding of how people value and perceive creative AI remains fairly limited, especially in the music domain. Indeed, the process and experience of engaging with music, both as a composer and a listener, involve cultural and emotional connections and aspects (e.g., rhythm, melody, harmony) that do not have direct parallels in other creative forms. In contrast, other art forms, such as painting or writing, allow for different forms of AI interaction and integration. Moreover, to date, research in the field of artificial intelligence and music has mostly centred around technology development (Civita et al., 2021) and copyright issues (Sturm et al., 2019). Regarding the former, Civita et al. (2021) identify a growing global interest in the field, with publications of commercial entities (e.g., Google, OpenAI, Amazon, Sony, Spotify) comprising the most cited works, thus underscoring the importance of this technology for the future of the music industry. As for the latter, using AI systems in music creation has elicited discussions regarding copyright and moral rights issues, presenting new challenges to intellectual property law in the music industry (Sturm et al., 2019). The debate primarily revolves around the level of human involvement in the creative process and using copyrighted music to train AI systems (Sturm et al., 2019).

A small body of literature has begun to move towards understanding AI's impact on music from a social and psychological stance (Civita et al., 2021; Hong et al., 2021; Novelli and Proksch, 2022). For instance, the recent introduction of the Expectancy Violation Theory (EVT) into the discussion (Hong et al., 2021) marks a significant shift towards examining human-computer interactions (HCI) within the musical domain. Specifically, EVT, originally relevant to interpersonal communication settings but evolved to the field of human-machine interactions, has been applied and validated to AI instrumental music, with expectations and their violations playing a role in the evaluation of classical AI music (Hong et al., 2021). How-

ever, while Hong et al. (2021) are among the first to make an academic contribution to this field, their study remains correlational, only involved two music genres, and, most importantly, may not be generalisable to vocal music, which is perceived differently from instrumental music (Roehm, 2001). This difference implies that EVT may operate differently depending on whether AI assumes the role of composer or singer. This research seeks to address these limitations by employing a research design that manipulates expectancy violations and causally estimates their influence on song evaluations, adding depth to EVT. Indeed, understanding the distinct outcomes of these roles will enrich EVT, offering insights into the 'why' and 'how' of the theory's varied applications.

In essence, the domain of AI-generated cloned vocals remains largely unexplored. While a handful of studies have delved into audio cloning within the media industry, particularly concerning deep fakes (Amezaga and Hajek, 2022), its application in the musical context has not been adequately addressed. This represents a significant research gap, partially due to the technology's early iterations being hindered by synthesised voices that sounded unnatural and robotic, lacking any personal touch (Story, 2019). Only with recent advances has the quality of speech synthesisers improved significantly (Amezaga and Hajek, 2022), prompting questions about the public's expectations and perceptions of songs produced with this technology and its potential to replace human singers (Coscarelli, 2023). As voice cloning technology approaches a level of development suitable for commercial use in the music industry, this study proposes to be the first to explore public reactions to its application in this context. We aim to leverage insights from broader discussions on creative AI applications and extend EVT to the evaluation of AI-cloned vocals in music. We seek to fill part of the existing research gap and propose an experimental study to deepen our understanding of human-machine interactions in the creative domain, investigating how people perceive and evaluate AI products in the creative industries differently and whether EVT applies equally to AI composers (Hong et al., 2021) as to AI singers.

## 1.3 Research Questions

Aside from legal and ethical considerations, songs featuring commercially available AI-cloned vocals may arrive soon. This prediction is reinforced when comparing the recent advancements in commercial AI instrumental music with the rapid progress of AI voice cloning technology, which now requires fewer audio samples to achieve higher accuracy (Arik et al., 2018). Yet, it remains unclear whether the public is ready for this innovation. Nonetheless, drawing parallels with AI instrumental music and leveraging insights from research in AI and creative fields, it becomes clear that people's attitudes and expectations will play a crucial role in determining whether AI-cloned vocals will set a new trend in music production and consumption. As a result, this research seeks to address the following research question:

**RQ1**: How do expectations about songs featuring AI-cloned vocals influence people's evaluation of such songs?

Moreover, because expectancy violations and confirmations can happen at any time, regardless of people's beliefs or attitudes about machines, EVT is anticipated to be a relevant framework through which to address the influence of expectations. Therefore, we also seek to address the following research question:

**RQ2**: How does violating expectations about songs with AI-cloned vocals influence people's evaluation of such songs?

## 1.4 Research Design Overview

We empirically address the research questions using an online survey experiment conducted among general music listeners residing in the European Union (EU). Focusing on this target population ensures the findings are relevant within the specific cultural dynamics and legal context of the EU, enhancing their applicability. On a practical level, the experiment aims to assess how individuals evaluate various song samples when exposed to different information about AI cloning technology and knowing that the samples feature AI-generated vocals.

Specifically, individuals are exposed to two contrasting types of information: one that positively frames AI-cloned vocals and one that negatively frames them. This strategy is designed to prime individuals' expectations towards either a positive or negative view before they listen to the samples. Additionally, the experimental design include two distinct types of samples: one featuring high-quality vocals and the other featuring low-quality vocals. This strategy has a twofold objective. First, it allows us to investigate how quality influences the evaluation of the samples. Second, since participants are primed with either positive or negative expectations, presenting them with varying vocal quality will accentuate the element of surprise (positive or negative) for some groups after listening to the samples – pivotal in determining the role of expectancy violations. By creating contrasting conditions, we effectively generate different groups in terms of levels of expectancy violations. Furthermore, the song samples span two distinct music genre – rock and pop – as well as both female and male vocals, thereby increasing the external validity of the findings across various genres and vocal types. Overall, this design facilitates the collection of data that can be rigorously analysed using regression analysis, ANCOVA, and Hayes' PROCESS to address the research questions.

## 1.5 Research Outline

The structure of the paper follows a logical progression. It begins with a review of relevant literature, which informs the development of five hypotheses. The methodology chapter then elaborates on the research design, the data collection procedures, and the techniques used for data analysis. Next, Chapter 4 presents the preliminary analysis conducted ahead of hypothesis testing. In Chapter 5, we proceed to test our hypotheses and present the results, which are then discussed in Chapter 6. Lastly, we address the study's limitations and provide suggestions for future research.

# 2 Literature Review

This chapter provides an overview of the literature relevant to our research question. The theories discussed in this chapter serve as the foundation for developing our hypotheses and conceptual model. The chapter begins with an overview of the evolution of voice cloning tech-

nology, from early text-to-speech systems to advanced AI applications in the music industry. Next, we review findings from previous studies on AI perceptions in the creative industry. We then introduce the Expectancy Violation Theory and discuss its applications in AI research. Following this, we address the role of music genres in music evaluations. Finally, the chapter concludes by presenting the hypotheses and conceptual model.

## 2.1 Voice Cloning: Origins and Evolution

Voice cloning originates from the development of speech synthesis and text-to-speech (TTS) technologies (Masood et al., 2023). The former, speech synthesis, entails the broader process of generating 'synthetic' speech using a computer or other machine. The latter, TTS, is a specific type of speech synthesis which aims to convert written text into speech (Taylor, 2009). The systems used for these purposes are known as speech synthesisers and can be developed as either software or hardware solutions. The performance of a speech synthesiser is determined by its naturalness – defined as how closely its output resembles the human voice – and its intelligibility – or how easily its output can be understood (Amezaga and Hajek, 2022). This chapter provides a brief historical overview of the evolution of speech synthesis and TTS, leading up to the development of voice cloning. It also addresses a second approach within speech synthesis known as speech-to-speech (STS) or voice conversion (VC), which voice cloning also leverages. Voice conversion, emerging in recent years, aims to transform a person's voice to sound like another's while preserving the linguistic content (Sisman et al., 2020). Lastly, the chapter discusses real-world applications of voice cloning and ethical considerations, particularly regarding the music industry, namely the primary context of interest for this research.

### 2.1.1 The Mechanical, Electrical, and Digital Era

The quest to create machines that could mimic human speech dates back hundreds of years (Benesty et al., 2008). The initial attempts in the late 18th and early 19th centuries were mechanical devices designed to simulate the human phonatory system (e.g., where sound is produced) and vocal tract. Notably, in 1780, physicist Christian Kratzenstein engineered the

first device capable of emitting the five vowel sounds – a, e, i, o, u. Shortly thereafter, in 1791, engineer Wolfgang von Kempelen introduced a mechanical device that could articulate speech at the word level. His invention, a more advanced mechanical simulation of human speech production, featured bellows, reeds, and pipes that could generate vowel and consonant sounds resembling a child's voice. The mechanical era continued throughout the 19th century with the development of other devices. Among these, Joseph Faber's talking machine stands out as the first to reproduce entire sentences (Story, 2019).

Then, in the early to mid-20th century, speech synthesis transitioned into the era of electrical systems. This transition began in 1922, when physicist John Q. Stewart developed the first fully electrical speech synthesis device, although limited to producing vowels and diphthongs (Steward, 1922; Story, 2019). Subsequently, in the 1930s, Homer Dudley introduced the VOCODER, short for voice encoder, a more advanced, fully electrical speech analyser and synthesiser. This technology was further refined into the VODER (voice operation demonstrator), which enabled speech generation through manual controls, including a keyboard, wrist bar, and foot pedals. Since the device's early demonstrations to the audience, Dudley showcased its potential for entertainment applications, including singing (Story, 2019). Nonetheless, the technology's first application was primarily pragmatic, focusing on the secure transmission of scrambled speech signals during World War II (Benesty et al., 2008).

In the late 1950s and 1960s, digital technology brought about a significant change in speech synthesis technology. Synthesisers were no longer required to be tangible, physical devices. Instead, algorithms with computational instructions could replicate the same processes. This transition from hardware-based synthesisers to software-driven algorithms led to significant improvements in terms of quality and efficiency, along with new applications for speech synthesis (Story, 2019). For instance, in 1961, John Larry Kelly leveraged an IBM 704 computer to 'sing' and recreate the song 'Daisy Bell'. Despite its distinctly robotic-sounding singing, this instance illustrates how far speech synthesis technology had advanced (Amezaga and Hajek, 2022). Nonetheless, the first commercially available text-to-speech system did not emerge until 1976, when Kurzweil introduced a reading machine designed to assist people with disabilities (Bell, 2023). Kurzweil's machine used concatenative synthesis, which stitches to-

gether small speech fragments from an extensive database to generate sound patterns (Story, 2019; Weitzman, 2022). Since then, the number of commercial speech synthesis products has increased, starting with early applications in the video game industry, as seen with the release of 'Stratovox' in 1980 (Weitzman, 2022), and extending to the integration into various computer operating systems throughout the 1980s, with the earliest being Apple Computer's MacInTalk in 1984 (Amezaga and Hajek, 2022).

However, even at this stage, the technology primarily generated robotic-sounding voices, the most famous example being 'Perfect Paul', Stephen Hawking's standard voice (Story, 2019). In the late 20th and early 21st centuries, to achieve voices that sounded more natural and human-like, research began to explore the use of artificial intelligence, leveraging deep learning techniques and neural networks (Weitzman, 2022). The ability of these methods to model complex patterns in data, thereby understanding context, intonation, and emotional cues, proved to be highly efficient (Hseih, 2024). As a result, the integration of these approaches led to the development of TTS systems capable of producing more natural and intelligible synthesised speech, paving the way for voice assistants such as Apple's Siri, introduced in 2010, and Amazon's Alexa, released in 2014 (Amezaga and Hajek, 2022).

### 2.1.2 Voice Cloning: Text-to-Speech

Voice cloning emerged in the late 2010s, pushing the boundaries of text-to-speech technology by aiming to recreate the unique voice of specific individuals. In particular, voice cloning extends beyond the generation of natural-sounding speech to include the replication of distinctive vocal characteristics of individual voices, such as tone, pitch, timbre, accent, and speaking style (Masood et al., 2023). To achieve this, the technology leverages advanced neural network architectures to analyse the characteristics of a target voice and train models capable of synthesising new speech that embodies these distinctive characteristics (Amezaga and Hajek, 2022). Among the notable breakthroughs, in 2016, DeepMind developed WaveNet, which employs a neural network designed to analyse and directly model raw waveforms, thereby mimicking any human voice with unprecedented realism (Masood et al., 2023; Van Den Oord et al., 2016). In 2017, researchers at Baidu released Deep Voice 3, a system that overhauled tradi-

tional TTS pipelines by replacing every component with advanced neural networks, achieving a new level of synthesis quality and efficiency (Ping et al., 2017). Additionally, in 2018, Google released Tacotron 2, an end-to-end TTS system that uses neural networks to produce highly natural and intelligible speech (Masood et al., 2023; Shen et al., 2018).

Figure 1, based on Masood et al. (2023), displays the typical pipeline of the latest TTS systems designed for voice cloning. The process begins with two primary inputs: text and audio samples of the target's voice. The text is processed by a text encoder, which converts it into a form the model can use. Meanwhile, the audio samples are processed by a speaker encoder, which uses neural network architectures to extract the unique characteristics and features of the target's voice. These encoded representations are then combined in a mapping stage, where the system aligns the text content with the voice features. Next, the combined data is passed through a decoder, which generates a raw audio signal based on the integrated information. Finally, a vocoder refines this raw signal into a coherent and natural-sounding speech, resulting in a new voice replicating the target voice's unique characteristics while conveying the desired textual content (Masood et al., 2023).

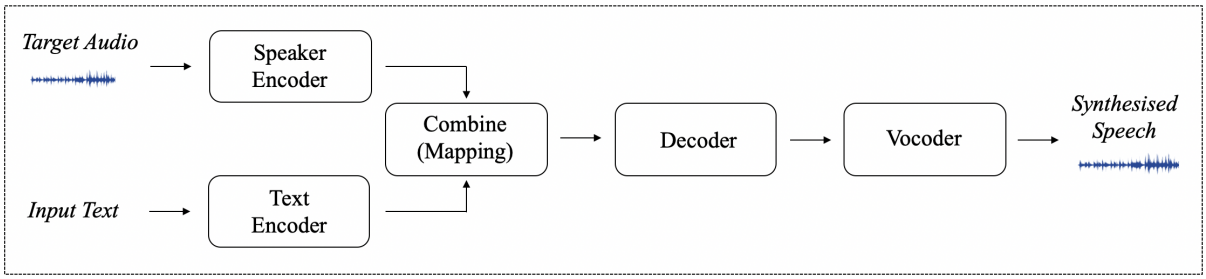



*Figure 1: The typical pipeline of the latest TTS systems*

The quantity of data (audio samples) required for voice cloning varies with the specific method used. A notable advancement in the field has been the significant reduction in the volume of raw data needed to produce convincing voice clones (Amezaga and Hajek, 2022). Presently, some models can generate convincingly natural speech and maintain a high degree of similarity to the original voice with just a few audio samples (Arik et al., 2018; Jia et al., 2018; Lee et al., 2018).

### 2.1.3 Voice Cloning: Speech-to-Speech

Within the domain of voice cloning, voice conversion has emerged alongside TTS as a speech-to-speech synthesis technique that transforms a speaker's voice to sound like another's without changing the content and structure of the spoken language (i.e., the linguistic information). Voice conversion thus differs from text-to-speech in that it directly transforms the properties of one voice to match another (Masood et al., 2023; Sisman et al., 2020). The technology began to take shape in the late 1980s and early 1990s, initially focusing on basic acoustic feature modifications to transform one speaker's voice (Machado and Queiroz, 2010). Over time, incorporating advanced neural network architectures has significantly improved the performance of VC systems, enabling more natural and accurate voice transformations (Masood et al., 2023; Sisman et al., 2020).

Figure 2, based on Walczyna and Piotrowski (2023), illustrates the typical pipeline of a voice conversion system designed for voice cloning. The process begins with two critical inputs: audio samples from the target voice and the source speech. For the target voice, in the features extraction phase, a speaker encoder captures its unique vocal characteristics, such as pitch, intonation, and acoustic features. Meanwhile, the source speech undergoes linguistic content extraction, which focuses on capturing its phonetic and prosodic information, such as its content, rhythm, and intonation. Hence, while the target features define how the converted speech should sound, the source content preserves what and how it is being said. The extracted target features and the source linguistic content are then processed by an encoder, combining and transforming them into a usable form. This encoded representation undergoes a mapping process where the speech features are adjusted to match the characteristics of the target voice, aligning the linguistic content with the target's vocal attributes. The mapped data is subsequently passed through a decoder, which generates a raw audio waveform based on the mapped information. Finally, a vocoder refines this raw waveform into a natural-sounding speech that maintains the target voice's unique characteristics while preserving the source speech's linguistic content (Sisman et al., 2020; Walczyna and Piotrowski, 2023).

The data requirements for voice conversion vary depending on the methods employed.

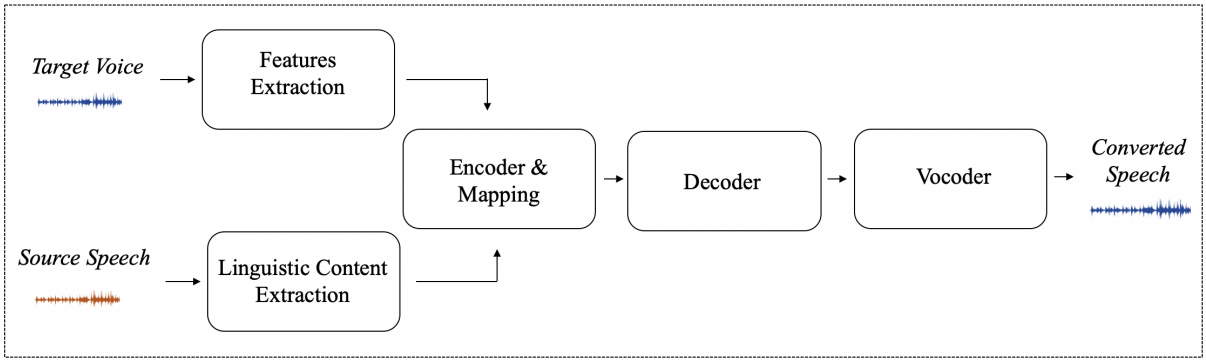



*Figure 2: The typical pipeline of a voice conversion system*

Earlier approaches necessitated the use of parallel data – datasets where the source and target speech are aligned with identical linguistic content. Although such data is often impractical for real-world applications (Masood et al., 2023), most literature on singing voice conversion has relied on parallel data (Nachmani and Wolf, 2019). Recently, there has been a shift towards methods using non-parallel data – the source and target speech do not need to be aligned – for both regular (Masood et al., 2023) and singing voice conversion (Chen et al., 2019; Nachmani and Wolf, 2019). Some newer approaches even involve cross-lingual voice conversion, where the source speech is transformed to sound like the target's using non-parallel data from different languages (Masood et al., 2023).

### 2.1.4 Real-World Applications

Voice cloning has seen many real-life applications across various industries. One prominent application is creating personalised conversational assistants, such as chatbots and AI assistants (Masood et al., 2023). For instance, in 2019, Cerence launched 'My Car, My Voice', allowing users to generate custom voices for their in-car assistant (Cerence, 2019). Voice cloning also offers personalised synthetic voices for vocally impaired individuals (Masood et al., 2023), helping those with conditions like motor neurone disease or Parkinson's maintain their vocal identity (Veaux et al., 2013). Additionally, text readers powered with voice cloning can be used to assist visually impaired individuals and non-readers and improve their educational experiences (Liu et al., 2015). Moreover, it can be used to translate videos and other content into multiple languages while preserving the original speaker's voice, allowing creators or businesses to reach a global audience (OpenAI, 2024). In the movie industry, it has been used for

automated dubbing (Sisman, 2020), replicating the voices of actors, and restoring or improving the quality of voices in older films (Vilenkin, 2023). However, its use has sparked ethical debates, as seen when it was used in Morgan Neville's 2021 documentary 'Roadrunner: A Film About Anthony Bourdain' to synthesise the voice of the deceased Bourdain (Hornaday, 2021). Additionally, recent protests from actors and voice artists have highlighted concerns about the unauthorised use of their voices (Shrivastava, 2023).

Voice cloning has also seen illicit applications, threatening voice biometric security systems (Masood et al., 2023), spreading fake news, and being used for political gains, identity theft and fraud (Wang et al., 2020). In particular, the emergence and growing popularity of audio deepfakes – synthetic audio recordings created using voice cloning to mimic a specific person's voice – pose significant challenges. These deepfakes undermine the credibility of auditory information and facilitate cyber-attacks, such as audio-based financial scams (Masood et al., 2023). For instance, in 2019, the CEO of a European company was deceived into transferring $243,000 to a scammer who cloned the voice of the head of the firm's parent company using publicly available audio recordings (Stupp, 2019). The potential use of such techniques to impersonate high-profile figures thus raises profound concerns, highlighting the dual-use nature of voice cloning technology (Amezaga and Hajek, 2022).

These and other concerns prompted OpenAI to delay the public release of Voice Engine, an advanced voice cloning tool. Specifically, OpenAI advocates for eliminating voice-based authentication for accessing bank accounts and sensitive data, establishing policies and regulations to safeguard individuals' voices, raising public awareness about the technology's capabilities and risks, and improving methods that detect and signal altered audio content (OpenAI, 2024). Nonetheless, other online platforms, such as Speechify or Respeecher, offer public access to voice cloning tools.

### 2.1.5 Applications in the Music Industry

Applying voice cloning in music production presents numerous opportunities for artists, music producers, and fans. First, it can facilitate sampling and remixing by allowing DJs or artists to generate new melody lines without requiring the original artist to record new vocals.

Second, it enables the production of songs using the cloned voices of deceased artists, either completing unfinished tracks or producing new ones. However, this requires careful ethical considerations and permissions. Third, it allows artists to venture into new genres without the need for extensive training. Fourth, it can also replicate the artist's voice at various pitches and tones, creating rich harmonies and eliminating the need for multiple vocal recordings. Furthermore, before completing a song, artists can use their cloned voices to test how specific lyrics sound and quickly make adjustments and improvements to the lyrics or melody. Finally, voice cloning allows artists or fans to easily create personalised songs (Weitzman, 2023b).

Recent studies have explored the application of voice synthesis and conversion systems in singing, employing various methods and achieving convincing results (Blaauw and Bonada, 2017; Chen et al., 2019; Choi et al., 2020; Lee et al., 2019; Nachmani and Wolf, 2019). These developments have fostered the creation of online tools for generating singing cloned vocals, such as Voice Swap, Kits AI, Triniti, and Jammable. Additionally, tools are available that do not strictly clone the vocals but rather imitate the styles of well-known artists, such as Jukebox by OpenAI (Dhariwal, 2020) and Suno. Nonetheless, their commercial applications are still limited by numerous ethical, privacy, and legal concerns that need to be addressed. For instance, some AI music generators are being misused to generate hateful songs (Wiggers, 2024). Additionally, there are concerns about the unauthorised use of an artist's voice and the need for explicit consent and compensation protocols (Weitzman, 2023b). For instance, in April 2024, Drake released a diss track titled 'Taylor Made Freestyle', which included cloned vocals of the deceased artist Tupac Shakur, without obtaining consent from Tupac's estate, resulting in the estate threatening legal action (Donahue, 2024).

In response to these challenges, several initiatives have aimed to explore the potential of voice cloning responsibly. For instance, in 2021, musician Holly Herndon released Holly+, a tool to create music using her vocal clone, allowing anyone to sing as 'Herndon', regardless of the language used. This tool can be used in both text-to-speech and speech-to-speech applications. All commercial decisions, control, and rights over Holly+ are managed through a decentralised autonomous organisation (DAO) based on blockchain technology. In 2019, Herndon also released PROTO, the first commercial album to use singing neural networks (Hern-

don, 2021). Another notable project is YouTube's DreamTrack, launched in 2023 (Cohen and Reid, 2023). Powered by Lyria, Google DeepMind's most advanced model (Google, 2023), the project involves nine famous artists and their vocals, including Charlie Puth, Demi Lovato, and John Legend. The experiment aims to explore the responsible use of AI music generation, including voice cloning, by artists and fans collecting feedback from a small group of artists and creators who have access to this technology to create YouTube Shorts (Cohen and Reid, 2023).

## 2.2 Public Perceptions of Creative AI

While prevailing studies on machine heuristics investigate machine-like performances of AI, such as news writing (Spence et al., 2019), a growing literature, whose objective is to understand machine's human-like traits (e.g., creativity), addresses creative machine heuristics to understand their capabilities and social roles (Hong et al., 2022). In this regard, Hong et al. (2022) explore how two traits of creative machine heuristics, namely autonomy and anthropomorphism, influence the evaluation of AI-composed music. They find that neither autonomy nor anthropomorphism plays a significant direct role in assessing AI-composed music. Hong et al. (2022) also find that an AI music generator with human-like traits is more likely to be accepted as a musician and its music to be appreciated, supporting role theory (Biddle, 1986) in that people's attitudes towards the role assigned to machines influence their performance assessment. These findings are supported by Sun et al. (2023), who reveal that higher anthropomorphism in AI musicians leads to increased perceptions of competence (i.e., intelligence, skill, creativity and efficacy), though not necessarily warmth (i.e., perceived friendliness, trustworthiness, kindness, and morality of an entity), and both dimensions positively affect attitudes towards AI as musicians. Furthermore, additional support is found by Messingschlager and Appel (2023), who explore how two dimensions of mind, namely experience (e.g., ability to feel emotions or having personality traits) and agency (e.g., ability to think or recognise emotions), influence the evaluation of AI artworks. They find that knowing an artwork was created by AI does not diminish its appreciation. However, human artists are attributed more experience and agency, which are linked to a positive evaluation of AI artworks (Messingschlager and Appel,

2023).

In regards to the general public's perception of creative AI, Moura and Maw (2021) find that while negative perceptions of AI musicians correlate with lower purchase intentions, the awareness of a composition's identity, whether human or AI, does not significantly influence the perception of music in terms of affective response, meaningfulness, and general attitude. Hong et al. (2021) explain these findings, suggesting that expectations and attitudes towards AI's creative abilities play a crucial role in how people value music. Therefore, in line with previous studies on AI artworks (Chamberlain et al., 2018; Hong and Curran, 2019), when people have faith in the creative abilities of AI, they appreciate its music more. This implies that if people consider singing a uniquely human trait, AI-cloned vocals might be undervalued simply because of their origin.

### 2.2.1 Hypotheses Development

Drawing on the insights and findings presented in Chapter 2.2, we intend to examine the specific case of AI-cloned song vocals. Given that singing is often perceived as a uniquely human trait, there may be additional layers of complexity in how the public evaluates AI-cloned vocals. Nevertheless, our first hypothesis is based on the idea that, in the creative industry, individuals with a favourable view of AI's technological capabilities are more likely to evaluate its products positively. This aligns with findings by Hong et al. (2021), Chamberlain et al. (2018), and Hong and Curran (2019), suggesting that belief in AI's creative abilities enhances the appreciation of its outputs. Additionally, the general acceptance of AI's musical capabilities and the lack of a significant impact of AI's identity on music perception (Moura and Maw, 2021) reinforce this hypothesis.

**H1:** The more positive the attitudes towards creative AI, the higher the evaluations of songs with AI-cloned vocals.

Our second hypothesis builds on role theory (Biddle, 1986), which posits that social roles are defined by the expected performances (or behaviours) associated with that role rather

than the performer's inherent characteristics. Applying this concept to AI-cloned vocals, if the public accepts AI as a legitimate participant in the role of a singer, their evaluation will be based on the performance of the song rather than the AI's non-human attributes. This implies that positive attitudes towards AI's role as a singer will likely lead to more favourable evaluations of songs featuring AI-cloned vocals. In the instrumental music domain, this is supported by findings from Hong et al. (2022), who show that acceptance of AI musicians leads to greater appreciation of their music, and Sun et al. (2023), who demonstrate that perceived competence of AI musicians results in higher acceptance.

**H2:** The higher the acceptance of AI as a singer, the higher the evaluations of songs with AI-cloned vocals.

## 2.3 Expectancy Violation Theory

Expectancy Violations Theory (EVT) explains and predicts how individuals respond to unexpected behaviours or actions in communication settings. It claims that people have expectations for how others should behave in a given context, originating from social norms, cultural norms, and individual idiosyncrasies (Burgoon, 1978). Within conversations and interactions, these expectations can either be confirmed or violated. The valence of the violation depends on whether one's expectations are exceeded (positive violations) or failed (negative violations). EVT argues that people will respond in specific and predictable ways when expectations are violated, either positively or negatively. Notably, positive violations lead people to perceive outcomes as more desirable than if the behaviours had conformed to expectations. In contrast, negative violations result in less favourable outcomes than if the expectations had simply been met (Burgoon, 2015). Whereas typical views often suggest avoiding all violations as inherently bad and leading to negative outcomes, EVT thus makes the counterintuitive argument that, if positive, violations can be beneficial (Burgoon, 2015).

Originally centred in the domain of interpersonal communication, EVT has evolved to include verbal, computer-mediated communications, and human-computer interactions (HCI),

extending its scope beyond personal interactions to different technological domains (Bonito et al., 1999; Edwards et al., 2016; Kalman and Sheizaf, 2010). For instance, Burgoon et al. (2016) extend EVT to interactions between humans and embodied agents (EAs) in decision-making tasks, while Waddel (2018) applies it to the media industry. The former study reveals that when EAs exceed expectations, they are viewed as more desirable than humans, whereas negative violations do not lead to less desirable perceptions. The latter study shows that negative expectancy violations towards alleged machine journalists lead to reduced news credibility.

EVT has also been extended to the creative industry in the art and music domain. Within the art domain, the concept of expectancy violation has been introduced by Messingschlager and Appel (2023) in their efforts to explain how the knowledge that an artwork is generated by AI can influence the appreciation of such artworks. Their research found a positive correlation between positive expectancy violations and artwork appreciation: the greater the positive deviation from expectations, the higher the appreciation of the artworks. However, their experiment's specific design and nature limit the broader implications of their findings on EVT, highlighting the need for further research in this domain.

Within the music domain, most studies focus on human-composed music, addressing the role of harmonic expectancy violations on neural responses (Janata, 1995) and musical emotions (Steinbeis et al., 2006). Hong et al. (2021) are the first to empirically observe expectancy violations in the context of AI and music, extending EVT to evaluations of an AI's performance in music composition. They observe that people's assessment of AI-composed classical music is influenced by the valence of their expectancy violations. In other words, whether the expectancy violation is positive or negative matters in how people value AI-composed classical music. Specifically, music that falls short of expectations receives lower evaluations compared to when it meets expectations; conversely, music that exceeds expectations receives higher evaluations than when expectations are confirmed. In contrast, they observe that for AI-composed EDM music, whether surprises in the quality of the music are perceived as positive or negative does not significantly influence people's evaluations of such music. This suggests that the effect of expectancy violations on AI-composed music evaluation may vary across different genres.

However, the study by Hong et al. (2021) has its limitations, which my research seeks

to address. First, the scope of their study is limited to two distinct music genres, classical and EDM, thereby limiting its generalisability. Additionally, the study overlooks the potential impact of people's genre-specific attitudes on music quality evaluations. Furthermore, the absence of experimental manipulations categorises their research as observational, precluding any inference of causality. Finally, introducing human or synthesised vocals may influence music evaluations and expectancy dynamics, as perceptions of instrumental and vocal music differ (Reid, 2001; Roehm, 2001).

While these studies have focused on how AI agents perform in specific domains (i.e., art and music), Hong (2021) adopts a broader perspective and considers whether the violation of expectations regarding AI's performances influences the adoption of overall AI technologies. Similar to what was observed by Hong et al. (2021), the only significant difference between groups with violated expectations and those with confirmed expectations is observed in negative, rather than positive, expectancy violations. In essence, negative surprises about what AI can do significantly reduce people's willingness to use AI technologies, while positive surprises do not increase the intention to adopt AI. However, the findings presented are correlational, as the study did not involve direct manipulations of expectations. Moreover, the survey overlooks the influence of participants' pre-existing knowledge and expertise in AI. These observations must be consistently replicated in future research across various contexts for a meaningful contribution to EVT.

### 2.3.1 Hypotheses Development

Based on the review of EVT and its application across AI and creative domains, this study aims to explore the impact of AI-cloned vocals on song evaluations. Our next hypothesis builds on the idea that expectancy is central to how individuals perceive and appreciate music (Hong et al., 2021). Expectations can be shaped by factors such as music genre (Istók et al., 2013; Squires, 2019) and external opinions (Jensen et al., 2013). Positive expectations about music quality can lead to more favourable evaluations, as listeners are primed to perceive the music positively (Hong et al., 2021). Thus, we posit that:

**H3:** Having positive expectations about the quality of AI-cloned vocals will lead to higher evaluations of songs with AI-cloned vocals compared to having negative expectations.

Moreover, higher quality outputs in music result in more favourable evaluations and greater acceptance (Schäfer and Sedlmeier, 2010), regardless of their origin (Hong et al., 2022). This implies that high-quality vocal performances may be generally perceived favourably, whether human or AI-generated. Thus, we further hypothesise:

**H4:** High-quality vocals will lead to higher evaluations of songs with AI-cloned vocals than low-quality vocals.

Lastly, EVT emphasises the greater impact of expectancy violations compared to expectancy confirmations. The theory indicates that unexpected events have a more pronounced effect than expected ones (Burgoon, 2015). Research shows that expectancy violations significantly influence AI product evaluations (Hong et al., 2021; Messingschlager and Appel, 2023) or adoption intentions (Hong, 2021). Specifically, expectancy violations in AI-generated instrumental music enhance appreciation and higher evaluations (Hong et al., 2021). Therefore, our fifth hypothesis proposes that the effect of expectations on the evaluation of songs featuring AI-cloned vocals will be stronger in the case of expectancy violations compared to expectancy confirmations.

**H5:** The effect of expectations on the evaluation of songs with AI-cloned vocals is stronger with greater levels of expectancy violations.

## 2.4 Music Genre

People have created labels to categorise and describe music. These labels, known as music genres, emerge from cultural dynamics and shared musical attributes, such as instrumentation, rhythmic structure, and melodic elements (Tzanetakis and Cook, 2002). Table 1 presents a short list of 14 genres identified in Rentfrow and Gosling (2003). In addition to

these broad genres, there are hundreds of subcategories that may intersect (Pachet and Cazaly, 2000), with numerous studies proposing different methods and taxonomies for their classification (Oramas et al., 2017).

*Table 1: List of 14 Music Genres*

| | |
|---|---|
| Alternative | Rap / Hip-Hop |
| Blues | Jazz |
| Classical | Pop |
| Country | Religious |
| Electronica / Dance | Rock |
| Folk | Soul / Funk |
| Heavy Metal | Sound Tracks |

This research specifically addresses pop and rock music. Pop music, short for popular music, is known for its simple beats, melodies, and harmonic elements and typically features repeated choruses and hooks (Steinbrecher, 2021). On the other hand, rock music revolves around musical instruments, especially bass guitar, drums, and electric guitar, and typically features more complex structures and lyrics (Covach, 1997). These two genres are often perceived differently in terms of their cultural status and accessibility. Pop songs are generally viewed as more mainstream and accessible, appealing to a wide audience through catchy hooks, melodies, and relatable themes (Steinbrecher, 2021). In contrast, rock songs often feature more intricate and less demanded themes (e.g., themes of rebellion or social issues), making them less accessible but more valued for their artistic expression and cultural significance (Regev, 1994). This framing may create a bias in people, likely influencing the evaluation of songs from either genre (Hong et al., 2021).

Previous research indicates that different music genres elicit distinct emotional responses, inducing a range of positive and negative emotions of varying intensities. For instance, rock music is linked to intense negative emotions such as anger and sadness and high energy levels, while pop music is generally associated with positive emotions such as happiness and joy (Rentfrow and Gosling, 2003). People develop stable preferences for music genres that persist throughout their lifetimes (Scherer and Zentner, 2001). These preferences originate from

personality traits (Rentfrow and Gosling, 2003), cultural trends, social interactions, and past experiences with music (Rentfrow et al., 2011). Typically, these preferences begin to form in early adolescence and solidify during late adolescence and early adulthood (Mulder et al., 2010). A study suggests that people's genre preferences and the emotional responses elicited by these preferences influence their evaluation and judgment of music (Istók et al., 2013). Therefore, when listening to songs with AI-cloned vocals, the music genre is likely to be a significant factor influencing their evaluation.

When applying expectancy violation theory, research indicates that music genres may influence how individuals evaluate music. Specifically, Hong et al. (2021) find that the violation of expectations regarding music quality affects the evaluation of AI-generated classical music. However, this effect is not observed for AI-generated EDM music. The authors suggest that the level of anthropomorphism in music genres and individuals' preferences may contribute to the interaction effect between the genre of AI music and expectations. However, they call for further research to investigate this relationship.

The use of classical and EDM music is appropriate for a study on instrumental music; however, the scope of this research calls for a different investigation, specifically on genres where vocals play a significant role. Therefore, we choose pop and rock music due to their differences from a musical perspective (Regev, 1994; Steinbrecher, 2021) and the contrasting personality traits associated with individuals favouring either genre (Rentfrow and Gosling, 2003). To date, no empirical study has addressed the evaluation of songs with AI-cloned voices or the influence of musical genres on such evaluations. Nonetheless, research suggests that people who favour rock music may be more curious and open to innovations (Rentfrow and Gosling, 2003). Hence, the applications of AI-cloned vocals – innovative technology in music production – might be perceived differently depending on the genre of the song and individual preferences, with rock enthusiasts potentially viewing it more favourably. Following this logic, to enhance our findings' generalisation and external validity, we aim to test our hypothesis across two distinct genres: pop and rock music.

## 2.5 Conceptual Model

Figure 3 visualises the conceptual model. The control variables included in the model are further explained in the methodology section, detailed in Chapter 3.6. An overview of the five hypotheses is provided in Chapter 5.3.

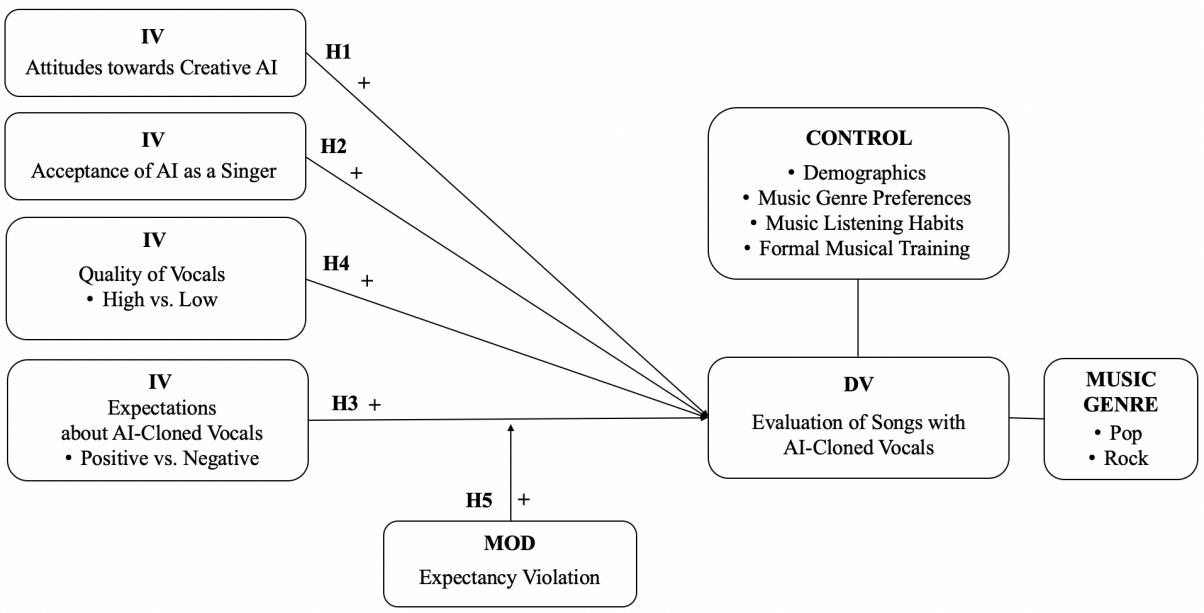



*Figure 3: Conceptual model*

# 3 Methodology

This chapter provides a detailed account of the methodology used to validate the conceptual model presented earlier (Figure 3). Starting with an explanation of the research design, the chapter delves into the experiment's core components. Finally, the chapter outlines the data analysis techniques applied in the study.

## 3.1 Research Design

Our research design follows a deductive approach, using theory to develop hypotheses, which are then validated using a suitable research method (Bhattacherjee, 2012). Since our study aims to determine cause-and-effect relationships, experimental research is the most appropriate. Indeed, experiments are highly valued for their internal validity and are typically regarded as the 'gold standard' of causal inference. Specifically, they provide a controlled

environment that minimises the impact of confounding factors (Bhattacherjee, 2012). By manipulating our independent and moderating variables, controlling for the effects of extraneous variables, and observing changes in our dependent variable, we can establish a clear direction of causality, thereby avoiding issues of reverse causality or simultaneity (Bhattacherjee, 2012).

### 3.1.1 Experimental Design

Our experiment requires the manipulation of three variables. Therefore, we adopt a 2x2x2 mixed factorial design consisting of three treatments, each with two levels. The independent variable, *expectations about AI-cloned vocals in songs*, is manipulated to be either positive or negative. The independent variable, *quality of AI-cloned vocals*, is manipulated to be either high or low. Since these two treatment conditions are not independent – that is, manipulating expectations may influence how subjects perceive the quality of the AI-cloned vocals – we employ a between-subjects design to ensure the effects of one treatment do not influence the other (Charness et al., 2012). The variable, *song genre*, has two levels: pop and rock. Since this variable is independent of the other variables, each level is assigned using a within-subjects design, meaning each subject in the experiment experiences both genres (Charness et al., 2012). Figure 4 provides an overview of the experimental design. In addition to these groups, we include two control groups where expectations are not manipulated, primarily for manipulation checks.

*Figure 4: Overview of the 2x2x2 factorial design*

| | | **Vocals Quality** | | | |
|---|---|---|---|---|---|
| | | Low | High | Low | High |
| **Expectations about AI Vocals** | Positive | Group 1 | Group 2 | Group 1 | Group 2 |
| | Negative | Group 3 | Group 4 | Group 3 | Group 4 |
| | | Pop | | Rock | |
| | | **Song Genre** | | | |

Since we aim for a true experimental design, subjects are randomly assigned to different treatment conditions (Bhattacherjee, 2012). Moreover, we employ counterbalancing by randomly assigning subjects to experience the song genres (pop and rock) in different orders. This approach mitigates order effects, including carryover, learning, and fatigue effects (Jhangiani et al., 2019).

## 3.2 Data Collection

To collect our experimental data, we designed a survey using Qualtrics. Qualtrics is a cloud-based platform typically used for creating surveys as well as collecting and managing survey data. It is well-suited for experiments, offering features such as random assignment to treatment conditions (Qualtrics, n.d.a), meta-information collection (Qualtrics, n.d.b), and timing tools which help control and determine the quality of survey submissions (Francis, 2014). The survey was distributed using Prolific, an online crowdsourcing platform specifically designed for academic and market research (Prolific, 2024a). Online crowdsourcing is a legitimate method to recruit participants for experimental studies (Casler et al., 2013). There is growing evidence that data collected via Prolific is comparable to that from laboratory studies (Palan and Schitter, 2018; Peer et al., 2017). Additionally, research shows that Prolific provides higher data quality compared to rival platforms and panels (e.g., Amazon Mechanical Turk, CloudResearch, Qualtrics, and Dynata). In particular, Prolific stands out in terms of participants' comprehension of instructions, attentiveness, and honesty in responses (Peer et al., 2022).

The data was collected between June 21st and 23rd, 2024. Participants were paid upon completing the survey, provided they passed two attention checks and completed the study on a desktop with working audio. The median completion time of the survey was 10:02 minutes for the treatment groups and 7:55 minutes for the control groups, resulting in an average reward per hour of £6.00/hr, in line with Prolific ethical payment principles (Denison, 2023). Additionally, the compensation rate of £6.00/hr is higher than typical rates (Prolific, n.d.). This is important, as Lovett et al. (2017) indicate that offering higher compensation results in higher data quality.

## 3.3 Experimental Procedure

We organised the study using a two-phase approach. First, we ran a pilot test with a smaller sample to determine whether all concepts and questions in the survey were clear and unambiguous, whether the random assignment was working correctly, and whether our manipulations were effective. We also sought validation of our survey design and experimental procedure from experts within the field of research methods. After making the necessary adjustments, we then distributed the final version of the survey on Prolific. This chapter details the final survey's design and flow, as depicted in Figure 6 preceding Chapter 4.

### 3.3.1 Pre-Survey Briefing

Before accessing the survey, all participants are presented with a pre-survey briefing, where we outline the survey tasks, estimated duration, and payment scheme upon successful completion. We also brief participants about the protocols for collecting, storing, and handling personal and survey data. Additionally, we warn them about the presence of attention checks, and that failure to appropriately respond to them will result in no compensation, as per Prolific policy (Prolific, 2024b). Moreover, we inform participants that they have the right to withdraw their participation at any time for any reason and that they would not be able to return to previously answered questions. Finally, after giving their voluntary consent to join the study, participants are invited to provide their Prolific ID and begin the survey. Appendix A reports the detailed pre-survey briefing as distributed on Prolific.

### 3.3.2 Baseline Measures

We initially define *vocals* and *AI-cloned vocals* to ensure that all participants are familiar with the study's relevant vocabulary (Appendix B). We then ask participants to report their views on creative AI and stances on AI taking the role of a singer. Additionally, we ask them about their music listening habits, previous formal musical training, and whether they have worked in the music industry. Participants are also asked to rate their preferences across 14 different music genres. Measuring these aspects prior to any experimental manipulations allows us to include them as covariates or controls within our study's analysis, thereby mitigating

the risk of bias that could arise from the experimental conditions influencing these responses (Field and Hole, 2002). Chapter 3.6 details how we operationalised and measured these constructs and motivates their inclusion in the experiment. Appendix C shows the format in which the survey presented these questions.

### 3.3.3 Expectations Manipulation

To manipulate our independent variable, namely expectations about AI-cloned vocals, we randomly assign participants one of two survey versions. In both versions, participants are asked to imagine themselves in a scenario where they come across an article in a respected magazine discussing AI technology that clones singers' voices. They are then asked to read this article. A scenario experiment is used because it allows us to create a controlled environment where participants' expectations can be influenced by the information presented (Gregory and Duran, 2001). The two articles are identical in structure but differ in their content. In one survey version, the article frames the technology as rudimentary and the quality of AI-cloned vocals as poor. In contrast, the other version presents the technology as highly advanced and the quality of AI-cloned vocals as remarkable. Both articles include fictional quotes from industry stakeholders and music fans. This approach helps prime participants' expectations, as research shows that reviews about a product influence perceptions towards it (Jensen et al., 2013). Both articles were entirely written by the researchers, a fact disclosed to the participants only upon completion of the survey. We also included a comprehension check to ensure participants carefully read and understood the article. Appendix D reports the two articles and comprehension questions presented to the participants.

### 3.3.4 Manipulation Check: Expectations

The primary purpose of providing different articles is to create two distinct levels of expectations: positive expectations in the group reading the positively framed article and negative expectations in the group reading the negatively framed article. To verify the success of our manipulation – whether the two groups have significantly different expectations after reading the articles – we then asked participants to report their expectations by answering a few ques-

tions. Chapter 3.6 details how we operationalised and measured expectations, and Appendix E shows the format in which the survey presented these questions.

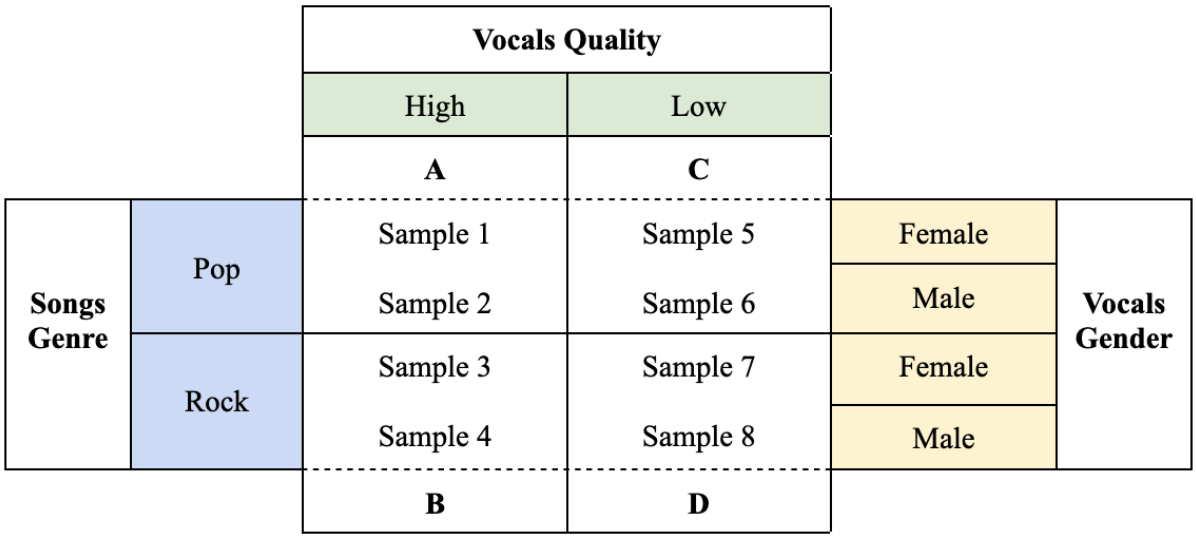
### 3.3.5 Vocals Quality & Music Genre Manipulations

After reading the article and reporting their expectations, participants are invited to listen to song samples which they believe feature AI-cloned vocals. We prepared eight different song samples: four featuring high-quality vocals and four featuring low-quality vocals. To manipulate the quality of the vocals, we selected real human songs for the high-quality samples and altered the vocals in these songs to sound more robotic and artificial for the low-quality samples. To avoid introducing geographic or cultural bias, all songs are performed in English, the most popular language for music across the EU (Spotify, 2024). Appendix H provides further details on the selection of the songs and the process of their manipulation. Within each quality category, we included two pop songs and two rock songs. To increase the generalizability and scope of the findings, we included one song by a male vocalist and one by a female vocalist in both the high-quality and low-quality samples for each genre. The samples are 15 seconds long, which may not fully represent how people listen to music in real-world settings. However, previous research shows that listeners can make stable and reliable evaluations in as little as 750 ms and that these evaluations are reinforced with longer exposure (Belfi et al., 2018).

First, we randomly assign participants two high-quality or low-quality samples from either the rock or pop genre. Figure 5 provides an overview of the four combinations of samples that each participant could be exposed to at this stage of the survey (i.e., A, B, C, or D). Upon listening to the two samples, we asked participants to assess how the AI-cloned vocals aligned with their expectations and to provide an overall evaluation of the samples. Chapter 3.6 addresses how we operationalised and measured these two constructs: our moderating and dependent variables. Appendix F shows the format in which the samples and subsequent questions were presented in the survey. Next, participants are given another pair of samples of the same quality as before but from the other genre. This means participants who initially received high-quality pop samples will now receive high-quality rock samples and vice versa. The same

applies to those who received low-quality samples. Participants are then asked again to compare the AI-cloned vocals to their expectations and to evaluate the samples. As mentioned in Chapter 3.1.1, we employ counterbalancing by randomly assigning participants to experience the song genres in different orders.

*Figure 5: Overview of the song samples matrix*




### 3.3.6 Manipulation Check: Vocals Quality

The primary purpose of manipulating the quality of the vocals in the samples is to create two distinct levels of quality: high and low. To verify the success of our manipulation – specifically, whether the manipulated vocals are perceived as having different quality – we collect data from two groups of participants who did not read any article before listening to the vocal samples and thus did not have their expectations manipulated. We then test whether the evaluation of the vocals differs between these control groups. Furthermore, having songs of varying vocal quality allows us to analyse expectancy violations more accurately (Olshavsky and Miller, 1972). In other words, this design accentuates the creation of different groups based on the interaction between expectations (positive or negative) and sample quality (high or low): those who are positively surprised by the quality of the vocals (negative and high), those who are negatively surprised (positive and low), or those whose expectations are met (negative and low or positive and high).

### 3.3.7 Demographics & Remarks

In the final survey component, we collect demographic information from the participants, including their gender, age, ethnicity, and education. These attributes are not affected by the treatments and do not demand focused attention; hence, they are placed at the end of the survey to prioritise our independent and dependent variables. Additional demographic details, such as country of residence and nationality, are provided by Prolific. Participants are also asked to detail any technical issues encountered during the survey, indicate whether they were familiar with the song samples, and provide final remarks. Finally, we disclose the true purpose of the study and the actual source of the articles and song samples. Appendix G includes this final block of questions as presented to the participants.

### 3.3.8 Attention Checks

To filter out inattentive participants, we include two attention checks to determine whether participants are paying attention and adequately following instructions. Specifically, we use instructional manipulation checks (IMCs) by incorporating two questions within the survey that are similar to the other survey questions in length and response format. The use of IMCs is associated with higher statistical power and data reliability (Oppenheimer et al., 2009) and is endorsed by Prolific (Prolific, 2024b). We place both attention checks at critical points in the survey: one within the items measuring the manipulation of expectations and the other within the items measuring the dependent variable related to the rock genre. Both checks draw on the same 7-point Likert scale; in the first case, we instruct participants to select 'Strongly disagree' (Appendix E), and in the second case, to select 'Strongly agree' (Appendix F).

### 3.3.9 Survey Design Elements to Mitigate Bias

Qualtrics allows for the inclusion of elements that help us minimise bias and collect valid and reliable data. First, we disable the option to return to previously answered questions to prevent participants from changing their responses based on later information, exposure to our manipulations, or perceived patterns in the survey. This is especially important when priming expectations and scenario-based questions, as the exposure to different information and

stimuli is controlled and sequential. Second, we include both positively and negatively worded questions to mitigate common method bias – which arises when data for both the independent and dependent variables are collected using the same method, leading to biased or inflated correlations (Kock et al., 2021). This approach helps reduce the likelihood of respondents consistently selecting the same position on a scale regardless of the question content. Moreover, we randomise how the items measuring each variable are presented to different participants. We also measure participants' time spent on each page to validate their responses. Third, differences in where and how participants listen to the song samples (environmental and technology bias) can affect their listening experience and evaluations. Therefore, we require all participants to take the survey on a desktop. To verify compliance, we collect meta-information at the start of the survey. Additionally, we recommend participants wear headphones throughout the survey. Although controlling for hardware is challenging, these guidelines aim to standardise the listening experience as much as possible. Finally, to further improve the validity and reliability of our data collection process, we implement additional elements such as forcing responses to all questions before allowing participants to move to the next page and hiding the 'Next Page' button for a reasonable amount of time depending on the content of the page.

## 3.4 Sample Size

The optimal sample size was determined using the G*Power software application, version 3.1.9.6 (Faul et al., 2007; Kang, 2021). To achieve sufficient statistical power, this study targeted a power level of at least 0.80 (Cohen, 1988). The recommended sample size was increased by 10% to account for invalid responses. A total of 14,196 individuals active on Prolific were invited to complete the survey. This participant pool included individuals from our target population who had at least five previous submissions on Prolific with a 100% approval rate. Setting these restrictions ensured participants were experienced with the platform and likely to provide high-quality data. A total of 229 participants voluntarily joined and completed the survey on Prolific.

## 3.5 Sample Population

The target population for our sample comprises individuals residing in the European Union (EU) who regularly listen to music. This population enables an investigation of policy implications unique to the EU, which has a specific regulatory framework for AI and digital technologies (European Commission, 2024). Moreover, the EU represents a major market for the music industry, with significant growth potential (European Commission, n.d.). Targeting general music listeners provides insights into this technology's acceptance and future marketability within this region. However, given the limited understanding of public perceptions of AI-cloned vocals in the music industry and the lack of empirical research on this topic, we decided to investigate a broad target population rather than focusing on specific demographics (e.g., age groups) or characteristics (e.g., musical preferences). This approach allows for a more comprehensive understanding of how AI-cloned vocals are perceived across different population segments, helps minimise bias that could arise from focusing on a narrow demographic, and lays the groundwork for more detailed future research.

We establish two primary criteria for validating survey submissions to ensure a representative sample population. First, only individuals residing in the European Union were invited to complete the survey. However, since the survey includes samples of songs performed by Italian artists, we excluded residents of Italy to reduce the likelihood of participants being familiar with these songs. Second, individuals were required to listen to music regularly, which was defined as at least one to two days per week. As a result, we excluded two participants who reported never listening to music.

We further validated survey submissions by inspecting several factors, including answers to attention checks, survey completion time, page timing statistics, meta-information, technical issues, and participants' familiarity with the song samples. One participant was excluded for failing one attention check and completing the survey on a mobile phone instead of a desktop. We estimated that going over the survey materials and answering the questions would take a minimum of 5 minutes and 15 seconds. Speeding (i.e., responding very quickly) can compromise response quality and is typically associated with straight-lining, where a participant repeatedly selects the same answer option across a series of questions (Zhang and

Conrad, 2014). Therefore, we excluded two submissions that were completed in less than 5 minutes. Additionally, two more participants were excluded for failing one attention check and completing the survey significantly faster than the average. Furthermore, two participants were excluded because they reported familiarity with the song samples, which might influence their evaluation. For each participant, we checked the time spent on each survey page, but no additional responses were excluded. Finally, one participant failed one attention check but mentioned that they '*[...] accidentally chose strongly disagree when asked to choose strongly agree'*; hence, we decided to keep their response. No participants reported technical issues with the survey.

## 3.6 Operationalisation of Constructs

Chapter 2 has defined the theoretical constructs and propositions of this research. To validate these propositions, we first operationalise our constructs by developing measurable indicators or metrics that accurately reflect them. This distinction is crucial because constructs exist at the theoretical level, whereas indicators (or variables, when combined) operate at the empirical level (Bhattacherjee, 2012). Instead of creating new indicators, which is beyond the scope of this research, we adapt existing measurement scales to fit the specificities of our research focus. Appendix I summarises and describes the adopted measurements of our constructs.

### 3.6.1 Dependent Variable

The song samples are evaluated using a revised version of a nine-item scale for assessing musical quality (Hickey, 1999). Hickey's (1999) scale has been a standard tool for evaluating musical quality, and its adoption in the context of AI-composed music by Hong et al. (2021) further supports its applicability to our research. The revisions made to the scale were carefully reviewed to fit the specificities of our research focus. All items in the scale are rated on a 7-point Likert-type scale ranging from 'strongly disagree' to 'strongly agree'.

### 3.6.2 Independent Variables

Expectations about AI-cloned vocals are manipulated to be either positive or negative by presenting participants with constructed information about the technology. Each participant is randomly assigned to read one of two narratives: the first narrative portrays AI-cloned vocals positively, whereas the second presents them negatively. The manipulation and content of the narratives are further discussed in Chapter 3.3.3. We also include a measurement of expectations to verify that the narratives effectively operationalise our construct by creating two specific levels of expectations among participants. This is done using a revised version of Olshavsky and Miller (1972), who, in their study, also created two conditions for expectations (positive and negative) by providing two different narratives and exposing participants to two levels of product performance (high and low). All items in the scale are rated on a 7-point Likert-type scale ranging from 'strongly disagree' to 'strongly agree'.

We include participants' acceptance towards AI taking on the role of singers, as previous studies have found it to influence product evaluations. Specifically, earlier research links positive evaluations of AI-composed music to the acceptance of AI as a musician (Hong et al., 2022), with further support in AI-generated artworks and the role of AI as an artist (Messingschlager and Appel, 2023). Additionally, the belief in AI's creativity is positively associated with evaluating AI-composed music (Hong et al., 2021). To measure participants' attitudes towards AI taking on the role of a singer, we ask them to report their agreement with three statements (e.g., 'I think the AI that clones vocals should be regarded as a singer'). These are evaluated on a 7-point Likert scale, where responses range from 'strongly disagree' to 'strongly agree,' with higher scores indicating greater acceptance of AI as a singer. The statements are adapted from a previous empirical study assessing the role of AI as a composer (Hong et al., 2022).

To measure participants' attitudes towards creative AI, we ask them to report their opinions on three statements (e.g., 'Products developed by AI should be respected as creative works') using a 7-point Likert scale ranging from 'strongly disagree' to 'strongly agree,' with higher scores indicating more positive attitudes. This measurement scale was used in a previous study on AI-composed music (Hong et al., 2021)

The quality of the vocals is manipulated to be either high or low by randomly presenting participants with either human songs or manipulated versions of those same songs. Additional information on the manipulation procedure is provided in Appendix H.

### 3.6.3 Moderating Variables

Expectancy violations and their valence are measured using a revised version of an instrument from a previous EVT study on AI-generated art (Messingschlager and Appel, 2023). Specifically, participants indicate to what extent the AI-cloned vocals deviate from their expectations using three items on a 7-point bipolar scale. This approach is chosen because it captures both the degree and direction of the violation, providing insights into whether expectations are violated and the valence of that violation. Initially, we tried to measure expectedness and valence separately with instruments from other EVT studies (Burgoon et al., 2016; Hong, 2021; Hong et al., 2021). However, these scales proved unreliable in our pilot test, leading us to adopt the bipolar measurements.

### 3.6.4 Control Variables & Music Genre

We include preferences for music genres as a control variable because previous studies have found that they play a crucial role in the evaluation and judgment of music (Istók et al., 2013; Shank et al., 2023). We measure participants' music genre preferences using the Short Test Of Music Preferences (STOMP) (Rentfrow and Gosling, 2003). The original STOMP test asks participants to rate their preference for 14 genres using a 7-point Likert-type scale ranging from 'dislike strongly' to 'like strongly'. Preference scores across these genres can then be aggregated to measure four factors: reflexive and complex (which includes the classical, jazz, and folk genres), intense and rebellious (which includes the rock and heavy metal genres), upbeat and conventional (which includes the pop, country, and soundtrack genres), and energetic and rhythmic (which includes the dance, rap, and soul genres). Each factor represents distinct dimensions of music preferences, providing a comprehensive overview of participants' musical tastes (Rentfrow and Gosling, 2003). A newer, revised version of the STOMP (STOMP-R) includes 23 genres, offering a more detailed and nuanced measurement of music preferences

(Rentfrow et al., 2011). However, to remain within the scope of this research and avoid overburdening participants, we find the original version, which also captures our genres of interest (pop and rock), to be adequate for this study.

We also include two control variables that measure individuals' engagement with music. Formal musical training is included as a control variable because previous studies indicate that music professionals exhibit different judgments to human music (Lehmann, 1997) and AI instrumental music (Shank et al., 2023) than non-musicians. Additionally, the idea of AI in music composition elicits different affective responses between musicians and non musicians (Moura and Maw, 2021). Another study found that individuals' listening habits significantly affect how they evaluate and experience music (Greasley and Lamont, 2011). Furthermore, long-term listening habits are associated with different emotional responses, influencing music evaluation (Rentfrow et al., 2011). Both variables are measured using elements from the Music USE (MUSE) questionnaire, an instrument developed to measure engagement in music (Chin and Rickard, 2012). Formal musical training is assessed with a multiple-choice question from Chin and Rickard's (2012) Index of Music Training (IMT), offering five options that range from 'less than one year' to 'more than ten years' of formal training. Listening habits are measured using Chin and Rickard's (2012) Index of Music Listening (IML), which aggregates scores from two multiple-choice questions. Both questions use a 5-point scale: the first assesses the frequency of weekly listening, ranging from 'never' to 'every day,' and the second measures the frequency of daily listening, ranging from 'less than one hour' to 'more than four hours.' The aggregated scores result in a value between 1 and 25.

Although random assignment should distribute demographic characteristics evenly across our groups, thereby helping to control for these variables, it relies on chance and may not achieve perfect balance (Bhattacherjee, 2012). Therefore, we collect this information and include it in our analysis as control measures if pre-analysis checks indicate an imbalance. Demographic information, including gender, age, ethnicity, and education, is measured using unidimensional scales, included in Appendix G.

To operationalise the music genre, we include rock and pop songs as two different levels. The songs were selected from artists typically performing in these genres. To ensure they

were recognisable and representative of their respective genres, we collected expert opinions that classified the songs as either pop or rock.

## 3.7 Data Analysis

We analyse the collected experimental data using a combination of Python, R, and SPSS. First, we conduct a preliminary analysis to test our scales' reliability and validity using Cronbach's alpha and Exploratory Factor Analysis, respectively. We then verify the success of our randomisation and experimental manipulations using parametric and non-parametric statistical tests, depending on whether the data follows a normal distribution. We also visually inspect our data with box plots, QQ plots, and histograms to detect outliers and distribution patterns, and lastly, we report descriptive statistics.

The hypotheses defined in Chapter 2 are tested using Ordinary Least Squares (OLS) regressions. OLS estimates the relationship between a continuous (or interval) dependent variable and one or more explanatory variables by minimising the sum of the squared differences between the observed and predicted values (Zdaniuk, 2014). This method provides coefficient estimates that indicate the strength and direction of the relationships. By examining these coefficients, we can determine whether the hypotheses are supported by the data.

To ensure our findings are robust across different statistical techniques, we further validate them using ANCOVA and Hayes' PROCESS macro. We decided to give precedence to OLS since it provides clear coefficient estimates that can be easily and directly interpreted to understand the strength and direction of relationships between variables (Zdaniuk, 2014). Additionally, we conduct both visual and statistical tests to validate the assumptions underlying our models.

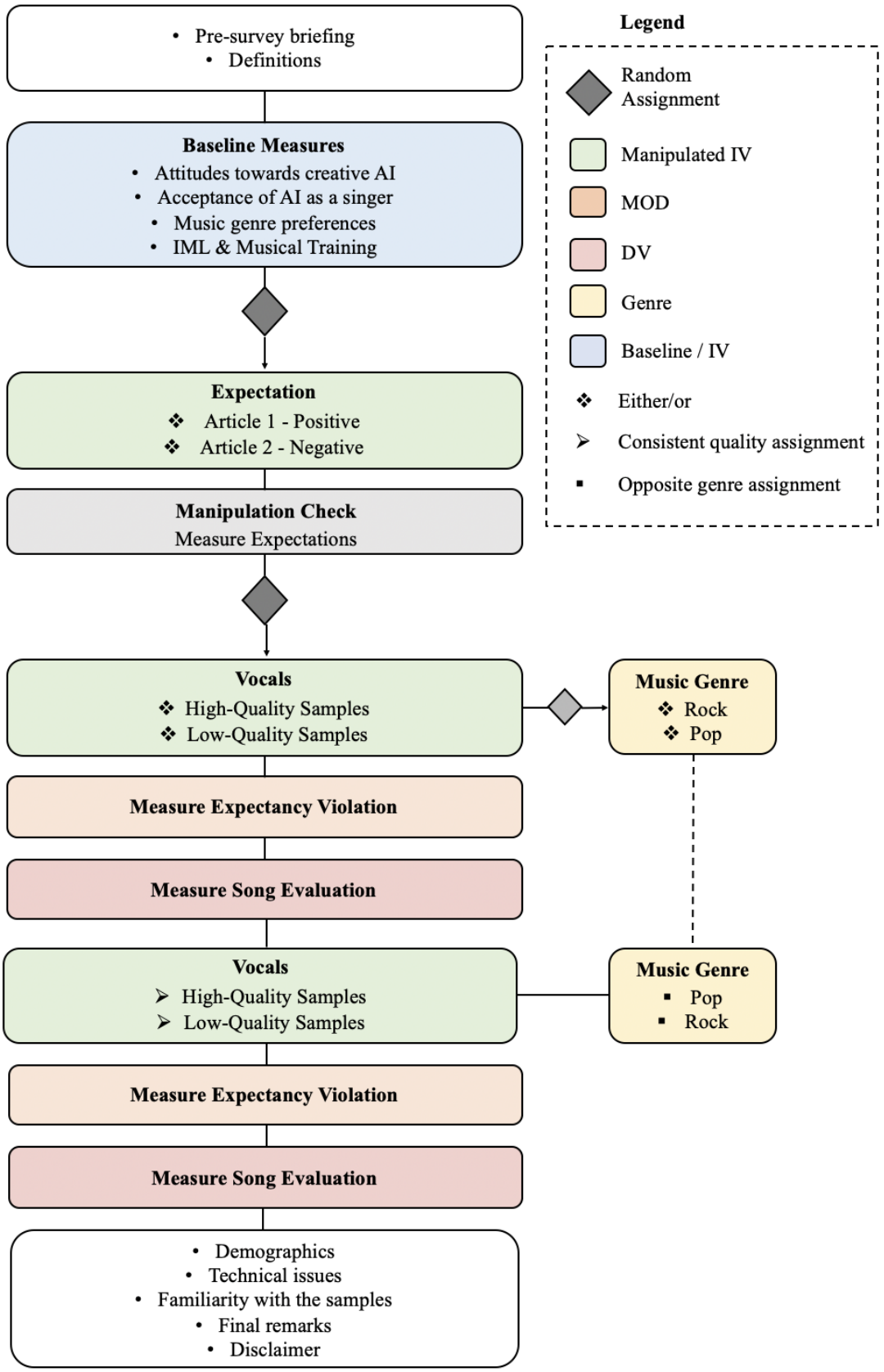



*Figure 6: Survey experimental flow*

# 4 Preliminary Analysis

This chapter elaborates on the preliminary analysis conducted before testing our hypotheses. First, we report our sample demographics. Then, since we use adapted scales from the literature, we test them to ensure they are valid and reliable. Specifically, we assess whether our multi-item measures adequately reflect the intended constructs (validity) and whether they do so consistently and precisely (reliability). Moreover, we verify whether our randomisation procedure has delivered comparable groups, safeguarding the study's internal and external validity. Then, we perform checks to verify the success of our manipulations, thus ensuring the causal validity of our experimental findings. Finally, we visually inspect our data and report descriptive statistics.

## 4.1 Sample Demographics

The sample demographics (Appendix J) indicate that the study population largely aligns with the target group, notwithstanding some minor discrepancies. First, all participants currently reside in the European Union, with 96% holding EU nationality. Additionally, the level of engagement with music is generally high, with 74% scoring at least a 6 on the index of music listening and 46% having at least one year of formal musical training. However, the gender distribution is predominantly male (58%), indicating an over-representation, while females represent only 41%. The age distribution highlights an under-representation of the older population, with only 8% of participants aged 45 and above. Additionally, the sample shows a high level of educational attainment, with over 71% achieving tertiary education (Bachelor's, Master's, professional, or Doctorate degrees). Ethnically, the sample is predominantly White (96%), with minimal representation from Asian (3%) and Hispanic or Latino (1%) backgrounds.

When we compare these demographics to the general population of the European Union, we notice some disparities. The EU has a more balanced gender distribution, with a nearly equal proportion of males and females (Eurostat, 2024b). Additionally, the EU's age structure is more evenly distributed across various age groups (Eurostat, 2024b), with a significant proportion of older adults reflecting an ageing population trend (Eurostat, 2024c). Regarding educational attainment, the broader EU population has a significantly lower proportion of

individuals with tertiary education (Eurostat, 2024a). The ethnic composition of the EU is also more diverse than the sample, with significant representation from various ethnic groups, including larger proportions of Asian and Hispanic populations (Eurostat, 2024c).

Some of these disparities are commonly observed when collecting data on Prolific and are regarded as the WEIRD bias (Prolific, 2023). Indeed, most participant pools in the social sciences skew towards more Western, Educated, Industrialised, Rich, and Democratic individuals (Henrich et al., 2010). The over-representation or under-representation of some groups poses limitations on the study's external validity, which we address in Chapter 7.

## 4.2 Scales Reliability

We estimate the reliability of our multi-item scales using Cronbach's alpha ($\alpha$), a measure of internal consistency designed by Lee Cronbach in 1951 (Cronbach, 1951). Cronbach's $\alpha$ estimates reliability in terms of average inter-item correlation, reflecting the extent to which items within a scale correlate and are consistent as a set (Bhattacherjee, 2012). The value of Cronbach's $\alpha$ ranges from 0 to 1, with higher values indicating higher internal consistency. Although the consensus on acceptable values and interpretations varies across research (Taber, 2018), a commonly accepted threshold for sufficient consistency is 0.70 or higher (Cho and Kim, 2015). Table 2 presents the Cronbach's $\alpha$ coefficients for our scales and their confidence intervals. All scales demonstrated significantly high internal consistency ($\alpha > 0.85$). Specifically, the scales measuring our independent variable reported a coefficient of 0.967 (95% CI: 0.960, 0.973) in the pop genre condition and 0.966 (95% CI: 0.960, 0.973) in the rock genre condition. While Cronbach's $\alpha$ generally increases with the number of items, and high scores may indicate that some items in the scale are redundant or overlap (Taber, 2018), subsequent analyses (e.g., evaluating the impact of removing items) revealed no need to drop any items. Therefore, we deem our scales suitable for further analysis.

*Table 2: Scales Reliability*

| | | 95% Confidence Interval | |
|---|---|---|---|
| **Scale** | **Cronbach's $\alpha$** | **Lower Bound** | **Upper Bound** |
| ATC | 0.878 | 0.846 | 0.904 |
| AS | 0.895 | 0.877 | 0.922 |
| E | 0.911 | 0.891 | 0.929 |
| EV Pop | 0.982 | 0.977 | 0.986 |
| EV Rock | 0.972 | 0.965 | 0.978 |
| SE Pop | 0.967 | 0.960 | 0.973 |
| SE Rock | 0.966 | 0.959 | 0.973 |

Note: Values are rounded to the nearest thousandth.

## 4.3 Scales Validity

We further validate our measures by conducting an Exploratory Factor Analysis (EFA) on our multi-item scales. EFA helps us understand the dimensionality of our scales and determine whether all items in each scale converge on the same construct (factor) and do not reflect unrelated constructs (Bhattacherjee, 2012). Following Hadi et al. (2016), as a first step, we assess whether our data is appropriate for EFA by testing sampling adequacy using Kaiser-Meyer-Olkin (KMO) (Kaiser, 1974) and the strength of the relationship among items using Bartlett's Test of Sphericity (Bartlett, 1950). Both tests confirm the data is suitable for further analysis. The results are reported in Appendix K; the KMO values are above 0.7, as recommended by Kaiser (1974), and Bartlett's Test of Sphericity reports a significant coefficient.

Following Samuel (2016), we began our EFA with Principal Components Analysis (PCA) to explore the underlying structure and estimate the number of factors. Then, we used Principal Axis Factoring (PAF) with oblimin rotation. The results, reported in Table 3, reveal a clear factor structure. Each item loads highly on only one factor, revealing five distinct constructs. For instance, items ATC1, ATC2, and ATC3 load highly on Factor 1, suggesting they measure the same underlying construct, whereas AS1, AS2, and AS3 load highly on Factor 2. The high loadings (generally above 0.7) indicate strong associations with their respective factors. Uniqueness values, representing the variance not explained by the factors, are generally low, supporting the adequacy of the factor model.

*Table 3: Exploratory Factor Analysis (Pop)*

| | Factor | | | | | |
|---|---|---|---|---|---|---|
| **Items** | **1** | **2** | **3** | **4** | **5** | **Uniqueness** |
| ATC1 | 0.842 | | | | | 0.249 |
| ATC2 | 0.735 | | | | | 0.401 |
| ATC3 | 0.896 | | | | | 0.211 |
| AS1 | | 0.907 | | | | 0.170 |
| AS2 | | 0.926 | | | | 0.166 |
| AS3 | | 0.781 | | | | 0.325 |
| E1 | | | 0.865 | | | 0.256 |
| E2 | | | 0.787 | | | 0.272 |
| E3 | | | 0.859 | | | 0.287 |
| E4 | | | 0.866 | | | 0.225 |
| EV1 | | | | 0.943 | | 0.045 |
| EV2 | | | | 0.963 | | 0.045 |
| EV3 | | | | 0.924 | | 0.062 |
| SE1 | | | | | 0.703 | 0.152 |
| SE2 | | | | | 0.943 | 0.197 |
| SE3 | | | | | 0.975 | 0.230 |
| SE4 | | | | | 0.813 | 0.077 |
| SE5 | | | | | 0.904 | 0.126 |
| SE6 | | | | | 0.736 | 0.158 |

Note: The 'Principal axis factoring' extraction method was combined with an 'oblimin' rotation. Values are rounded to the nearest thousandth.

These results provide strong evidence for the validity of our measurement scales, confirming that moving forward with our analysis is appropriate. Indeed, all items in our scales converge on their intended constructs without overlapping. For example, items intended to measure attitudes towards creative AI (i.e., ATC1, ATC2, and ATC3) converge on a single factor (Factor 1), and items intended to measure our independent variable (i.e., EV1, EV2, EV3, EV4, EV5, and EV6) also load distinctly on their respective factor.

We also report the inter-factor correlation matrix (Appendix K), which indicates generally low correlations among our factors, except for Factor 1 and Factor 2 (0.36) and Factor 4 and Factor 5 (0.80). Nonetheless, this aligns with our theoretical expectations. Indeed, attitudes towards creative AI and views on the role of AI as a singer are related (i.e., they both concern

the impact of AI on the music industry) but distinct constructs. Similarly, expectancy violations and evaluations of songs are related (i.e., they both involve listeners' perception and reaction to the songs) yet distinct constructs. The results of an equivalent EFA conducted with the rock condition are included in Appendix K.

## 4.4 Randomisation Check

Random assignment serves two crucial purposes (Bhattacherjee, 2012). First, it ensures internal validity by distributing control characteristics evenly across groups, thereby cancelling their extraneous effects. Second, it supports external validity by allowing the generalisation of inferences drawn from the sample to the population from which it is drawn (Bhattacherjee, 2012). However, our random assignment procedure relies on chance and may not have achieved perfect balance. For this reason, we tested whether the groups receiving different articles and those receiving different-quality songs differ in their characteristics. The choice of a factor-wise comparison – comparing the groups receiving different levels of each manipulated factor – simplifies the analysis while still providing robust information about the success of the randomisation, reducing the risk of Type I errors (false positives) that can occur when conducting multiple comparisons (Field et al., 2012).

To choose the appropriate test for our independent groups, we first checked whether the data was normally distributed and whether the assumption of homogeneity of variances was satisfied. The results of these tests are reported in Appendix L. Because the data is not normally distributed, we decided to run the non-parametric Mann-Whitney U Test for our ordinal and continuous variables (McKnight and Najab, 2010). As for the categorical variables, we opted for the Chi-Squared test of independence (Bhattacherjee, 2012). Results are presented in Tables 4 and 5, which shows no significant differences across groups, giving us reasonable confidence that the randomisation was successful.

*Table 4: Random assignment Check - Mann-Whitney U Test*

| | Article 1 vs. Article 2 | | High vs. Low Quality | |
|---|---|---|---|---|
| **Statistic** | **Statistic** | **p** | **Statistic** | **p** |
| Age | 3102 | 0.951 | 3036 | 0.770 |
| IML | 2979 | 0.622 | 3086 | 0.905 |
| Musical Training | 2992 | 0.625 | 3093 | 0.918 |
| Music Industry | 3038 | 0.392 | 3044 | 0.428 |
| Reflective Complex | 3089 | 0.915 | 2948 | 0.550 |
| Intense Rebellious | 3037 | 0.772 | 2770 | 0.222 |
| Upbeat Conventional | 3037 | 0.773 | 3032 | 0.758 |
| Energetic Rhythmic | 2924 | 0.493 | 2910 | 0.464 |

Note: Ha $\mu_0 \neq \mu_1$

*Table 5: Random assignment Check - Chi-Squared Test*

| | Article 1 vs. Article 2 | | | High vs. Low Quality | | |
|---|---|---|---|---|---|---|
| **Statistic** | $\chi^2$ | **df** | **p** | $\chi^2$ | **df** | **p** |
| Gender | 0.83 | 2 | 0.659 | 4.27 | 2 | 0.118 |
| Nationality | 24.1 | 26 | 0.571 | 31.1 | 26 | 0.225 |
| Education | 2.18 | 2 | 0.336 | 0.374 | 5 | 0.996 |
| Ethnicity | 2.33 | 5 | 0.801 | 0.234 | 2 | 0.889 |

Note: Ha $\mu_0 \neq \mu_1$

## 4.5 Manipulation Checks

After assessing the data's normality and homogeneity of variance (Appendix M), we conducted three independent Student's t-tests to compare the means across the three conditions. The results in Tables 6, 7, and 8 indicate significant differences in expectations ($p < .001$), confirming the success of our manipulation. Specifically, the group receiving Article 1 had significantly higher expectations, on average, than both the group receiving Article 2 ($t = 8.43$, $p < .001$) and the control group ($t = 3.40$, $p < .001$). Additionally, on average, participants receiving the negatively framed article showed significantly lower expectations than the control group ($t = -4.52$, $p < .001$).

*Table 6: Manipulation Check - Expectations: Article 1 vs. Article 2*

| | Article 1 Group | | | Article 2 Group | | | Student's t | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Expectations | 72 | 4.89 | 1.30 | 79 | 3.11 | 1.28 | 8.43 | <.001 |

Note: Ha $\mu_0 \neq \mu_1$

*Table 7: Manipulation Check - Expectations: Article 1 vs. Control*

| | Article 1 Group | | | Control Group | | | Student's t | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Expectations | 72 | 4.89 | 1.30 | 62 | 4.11 | 1.33 | 3.40 | <.001 |

Note: Ha $\mu_0 \neq \mu_1$

*Table 8: Manipulation Check - Expectations: Article 2 vs. Control*

| | Article 2 Group | | | Control Group | | | Student's t | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Expectations | 79 | 3.11 | 1.28 | 62 | 4.11 | 1.33 | -4.52 | <.001 |

Note: Ha $\mu_0 \neq \mu_1$

We also expose participants to two different quality levels of song samples: high-quality with human vocals and low-quality with poor vocals. To verify that the vocals differ in quality, we compared participants' vocal evaluations in the control condition. The control groups were not exposed to any priming articles, allowing their evaluations to serve as a baseline for attributing differences to the vocals' quality. After assessing the data's normality and homogeneity of variance (Appendix M), we conducted two Mann-Whitney U tests, one for the pop vocals and the other for the rock vocals. The results in Tables 9 and 10 indicate significant differences in the evaluation of the vocals ($p < .001$), confirming the success of our manipulation. Specifically, the group receiving the high-quality vocals had higher evaluations for both the pop ($U = 41.5$, $p < .001$) and rock samples ($U = 77.0$, $p < .001$), as indicated by the higher mean ranks.

*Table 9: Manipulation Check - Pop - High vs. Low Quality*

| | Low-Quality Group | | | High-Quality Group | | | Mann-Whitney U | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Songs Evaluation | 30 | 2.04 | 1.13 | 32 | 5.26 | 1.10 | 41.5 | <.001 |

Note: Ha $\mu_0 \neq \mu_1$

*Table 10: Manipulation Check - Rock - High vs. Low Quality*

| | Low-Quality Group | | | High-Quality Group | | | Mann-Whitney U | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Songs Evaluation | 30 | 2.54 | 1.48 | 32 | 5.26 | 1.14 | 77.0 | <.001 |

Note: Ha $\mu_0 \neq \mu_1$

Additionally, we examined whether the combination of receiving a specific article and song samples successfully created different expectancy violations among the groups. Specifically, we aimed to understand whether the groups receiving Article 1 had their expectations more violated when listening to the low-quality songs compared to those receiving Article 2, and vice versa for the high-quality songs. After assessing the data's normality and homogeneity of variance (Appendix M), we conducted separate Mann-Whitney U tests for the high-quality and low-quality groups. The results in Tables 11 and 12 indicate the successful outcome of our experimental design. Specifically, the group exposed to the negative article (Article 2) receiving the high-quality samples experienced a significantly greater expectancy violation for both the pop (U = 368, p = 0.041) and rock samples (U = 430, p = 0.045) compared to the group with the positive article (Article 1) receiving the same high-quality songs. Similarly, the groups receiving the low-quality samples showed significantly different violations for both the pop (U = 456, p = 0.012) and rock samples (U = 337, p < 0.001), as indicated by the higher mean ranks of the groups receiving Article 2. Thus, we can proceed with further analysis, knowing that our manipulation was successful in creating different expectancy violations among groups 1 and 3, as well as groups 2 and 4 (refer to Figure 4).

*Table 11: Manipulation Check - Expectancy violation high-quality samples*

| | Group 2 | | | Group 4 | | | Mann-Whitney U | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Violation Pop | 40 | 1.32 | 1.04 | 36 | 1.76 | 0.81 | 368 | 0.041 |
| Violation Rock | 40 | 1.15 | 1.21 | 36 | 1.70 | 1.14 | 430 | 0.045 |

Note: Ha $\mu_0 \neq \mu_1$

*Table 12: Manipulation Check - Expectancy violation low-quality samples*

| | Group 1 | | | Group 3 | | | Mann-Whitney U | |
|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **N** | **Mean** | **SD** | **Statistic** | **p** |
| Violation Pop | 32 | -2.24 | 0.77 | 43 | -1.47 | 1.29 | 456 | 0.012 |
| Violation Rock | 32 | -1.90 | 0.78 | 43 | -0.81 | 1.26 | 337 | <.001 |

Note: Ha $\mu_0 \neq \mu_1$

## 4.6 Visual Inspection

We inspected the data using box plots for outliers and distribution patterns (Appendix N). We found two outliers for our independent variable and four for our moderating variable, represented by the dots in the box plots. There is no single agreed-upon method for addressing outliers, and opinions on the best approach vary widely (Osborne and Overbay, 2019). Upon careful inspection and considering our models' sensitivity to outliers, we excluded them from further analysis to ensure the robustness of our statistical results. Additionally, the box plots indicate that the low-quality condition (Groups 1 and 3) exhibits more variability, with less consistent and more dispersed evaluations and violations in both genres, suggesting that the quality factor may significantly contribute to people's reactions. We further inspected our dependent variable using distribution histograms. Overall, the histograms suggest a non-normal distribution in one condition for our independent variable. This is further supported by the analysis using QQ plots and the values of skewness and kurtosis. These findings prompt us to further investigate the distribution of our data with statistical tests before testing our hypotheses, as a non-normal distribution may influence the results of our models.

## 4.7 Descriptive Statistics

Tables 13 and 14 present descriptive statistics for our dependent and moderating variable for each treatment condition, providing insights into the count, mean, standard deviation, minimum, and maximum values. Inspecting the counts, we observe that although we applied random assignment to treatment conditions, excluding some observations for poor response quality resulted in slightly uneven distributions. Nevertheless, the sample sizes per treatment condition remain above 30 observations, which should minimise potential consequences on the study's ability to identify significant differences between the groups.

Inspecting the means of our dependent variable (Table 13), in both genres (pop and rock), participants who received the positive article consistently evaluated the songs higher than those who received the negative article, regardless of the song quality. This may suggest that being presented the positive article has positively affected the participants' evaluations of the songs. Additionally, the difference in means seems to be more pronounced for the high-quality songs than the low-quality songs, indicating that the positive article may have had a stronger impact when the song quality was high. The descriptive statistics for expectancy violations (Table 14) indicate that participants who received the positive article experienced, on average, greater violations with low-quality songs and lower violations with high-quality songs across both pop and rock genres than those who received the negative article. This may suggest that the positive article has influenced participants' expectations, leading to greater disappointment with low-quality songs and lesser disappointment with high-quality songs.

*Table 13: Song Samples Evaluation*

| | **Evaluation Pop** | | | | | **Evaluation Rock** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **Min** | **Max** | **N** | **Mean** | **SD** | **Min** | **Max** |
| Overall | 151 | 3.62 | 1.73 | 1.00 | 7.00 | 151 | 3.80 | 1.65 | 1.00 | 7.00 |
| Group 1 | 32 | 2.34 | 0.99 | 1.00 | 5.17 | 32 | 2.60 | 0.96 | 1.00 | 4.83 |
| Group 2 | 40 | 5.14 | 1.09 | 2.17 | 7.00 | 41 | 5.13 | 1.12 | 2.50 | 7.00 |
| Group 3 | 43 | 2.10 | 0.93 | 1.00 | 4.83 | 43 | 2.51 | 1.07 | 1.00 | 4.67 |
| Group 4 | 36 | 4.89 | 1.02 | 2.50 | 7.00 | 37 | 4.92 | 1.27 | 2.50 | 7.00 |

Note: Values are rounded to the nearest cent.

*Table 14: Descriptive Statistics Expectancy Violation*

| | Expectancy Violation Pop | | | | | Expectancy Violation Rock | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **N** | **Mean** | **SD** | **Min** | **Max** | **N** | **Mean** | **SD** | **Min** | **Max** |
| Overall | 151 | -0.12 | 1.99 | -3.00 | 3.00 | 151 | -0.03 | 1.73 | -3.00 | 3.00 |
| Group 1 | 32 | -2.24 | 0.77 | -3.00 | 0.00 | 32 | -1.90 | 0.78 | -3.00 | 0.00 |
| Group 2 | 40 | 1.32 | 1.14 | -0.33 | 3.00 | 41 | 1.07 | 1.29 | -1.00 | 3.00 |
| Group 3 | 43 | -1.47 | 1.29 | -3.00 | 2.00 | 43 | -0.81 | 1.26 | -3.00 | 1.00 |
| Group 4 | 36 | 1.76 | 0.81 | 0.00 | 3.00 | 36 | 1.35 | 1.15 | -1.00 | 3.00 |

Note: Values are rounded to the nearest cent.

We also report descriptive statistics for our covariates and control variables in Table 15. Our sample shows generally low acceptance of AI as a singer, with a mean score of 1.90. This implies that participants are generally reluctant to the idea of AI replacing human singers. In contrast, attitudes towards creative AI are more positive and varied, as indicated by the mean of 3.29 and standard deviation of 1.47. Thus, while AI as a singer is viewed negatively, the more general concept of AI in the creative industry is accepted and regarded more positively. This aligns with previous research, finding that people are willing to accept and welcome AI artworks (Hong and Curran, 2019) and AI instrumental music (Moura and Maw, 2021). It might be that singers are viewed even more strongly as occupying a uniquely human-reserved role compared to other art forms.

The mean scores for different music genres vary, with rock (5.78), pop (5.50), and soundtracks or theme songs (5.69) being the most preferred genres, while Religious music has the lowest mean score (2.53). While the high mean score for soundtracks may come as a surprise, the rock and pop genres were selected for this study due to their popularity among residents of the European Union (Spotify, 2024). Hence, this choice is further corroborated by these results. Regarding the overall music dimensions, the 'Intense and Rebellious' dimension (which includes rock and heavy metal genres) seems to have greater variability, with the lowest minimum values and the highest maximum values among the four dimensions. This is also in line with previous research indicating that these dimensions elicit stronger emotions compared to the other three, thus describing a more love-hate dimension (Rentfrow and Gosling, 2003).

*Table 15: Descriptive Statistics Control & Covariates*

| | **Mean** | **Median** | **SD** | **Min** | **Max** |
|---|---|---|---|---|---|
| IML | 11.50 | 10.00 | 7.27 | 2.00 | 25.00 |
| Musical training | 1.85 | 1.00 | 1.19 | 1.00 | 5.00 |
| Attitudes Creative AI | 3.29 | 3.33 | 1.47 | 1.00 | 7.00 |
| Acceptance of AI as a singer | 1.90 | 2.00 | 0.98 | 1.00 | 5.67 |
| Classical | 5.03 | 5.00 | 1.53 | 1.00 | 7.00 |
| Blues | 4.67 | 5.00 | 1.32 | 1.00 | 7.00 |
| Country | 3.74 | 4.00 | 1.53 | 1.00 | 7.00 |
| Dance or electronica | 5.03 | 5.00 | 1.52 | 1.00 | 7.00 |
| Folk | 4.15 | 4.00 | 1.42 | 1.00 | 7.00 |
| Rap or hip hop | 4.87 | 5.00 | 1.65 | 1.00 | 7.00 |
| Soul or funk | 4.48 | 5.00 | 1.42 | 1.00 | 7.00 |
| Religious | 2.53 | 2.00 | 1.31 | 1.00 | 5.00 |
| Alternative | 5.03 | 5.00 | 1.44 | 1.00 | 7.00 |
| Jazz | 4.68 | 5.00 | 1.56 | 1.00 | 7.00 |
| Rock | 5.78 | 6.00 | 1.31 | 2.00 | 7.00 |
| Pop | 5.50 | 6.00 | 1.45 | 1.00 | 7.00 |
| Heavy metal | 4.17 | 4.00 | 2.03 | 1.00 | 7.00 |
| Soundtracks or theme song | 5.69 | 6.00 | 1.01 | 2.00 | 7.00 |
| **Genre Dimensions** | | | | | |
| Reflective complex | 4.63 | 4.75 | 1.08 | 1.75 | 6.75 |
| Intense rebellious | 5.00 | 5.33 | 1.20 | 1.33 | 7.00 |
| Upbeat conventional | 4.37 | 4.50 | 0.82 | 2.00 | 6.25 |
| Energetic rhythmic | 4.79 | 5.00 | 0.99 | 2.00 | 7.00 |

Note: N = 151. Values are rounded to the nearest cent.

# 5 Results

This chapter proceeds to test the hypotheses and report the results. Drawing on previous research by Hong et al. (2021) and Hong (2022), we first determine whether a linear relationship exists between two independent variables – *attitudes towards creative AI* and *acceptance of AI as a singer* – and the dependent variable - *song evaluation* - and whether these should be considered for further testing. Additionally, we report two robustness checks that corroborate our findings. Based on the simple linear regression results, *attitudes towards creative AI* and *acceptance of AI as a singer* are included as covariates in subsequent analyses for further hypotheses testing. Then, all hypotheses are tested using multiple regression analysis. Although additional control variables were deemed unfit for inclusion in the main model due to the lack of a linear relationship with the dependent variable, we report a second model incorporating them as a robustness check in Appendix Q. Additionally, an ANCOVA is conducted to provide an additional layer of robustness. Furthermore, we use Hayes' PROCESS macro for SPSS (Hayes and Rockwood, 2017) to provide additional evidence for the absence of a moderation effect. The chapter also includes a detailed examination of all the models' assumptions before presenting the results.

## 5.1 Preliminary Testing with Simple Linear Regressions

Following Hong et al. (2021) and Hong (2022), we use simple linear regressions to determine whether attitudes towards creative AI, hereinafter referred to as *Attitudes_creative_AI*, and acceptance of AI as a singer, hereinafter referred to as *Acceptance_AI_singer*, should be included in the multiple regression for further testing. This method quantifies the strength and direction of the relationship between an independent and a dependent variable (Field et al., 2012). Specifically, we run four regression models (Table 16) on evaluations of pop and rock music. Since our models display dependent errors, we fit the simple regression models with robust standard errors (HAC3). Additionally, we provide two robustness checks using bootstrapping and Spearman's correlation coefficients due to the non-normally distributed errors. These checks ensure our findings' reliability and validity despite the violations of regression assumptions.

*Table 16: Simple Linear Regression Models*

| **Model** | **Regression Equation** |
|---|---|
| Model 1 - Pop | $Song\_Evaluation = b_0 + b_1 \times attitudes\_creative\_AI + \varepsilon_i$ |
| Model 2 - Rock | $Song\_Evaluation = b_0 + b_1 \times attitudes\_creative\_AI + \varepsilon_i$ |
| Model 3 - Pop | $Song\_Evaluation = b_0 + b_1 \times acceptance\_AI\_as\_singer + \varepsilon_i$ |
| Model 4 - Rock | $Song\_Evaluation = b_0 + b_1 \times acceptance\_AI\_as\_singer + \varepsilon_i$ |

### 5.1.1 Assumptions: Simple Linear Regression

Simple linear regressions require the data to comply with four primary assumptions: a linear relationship between the independent and dependent variables, independence of errors, homoscedasticity, and normally distributed errors (Field et al., 2012). Therefore, we test each assumption for each regression model before conducting further analysis. First, we verify the linearity between the dependent and independent variables using scatter plots. The scatter plots for both genres show a clear linear relationship, satisfying the linearity assumption.

Second, we plot the residuals against the fitted values of the regression models to visually inspect the independence of errors and homoscedasticity. The residuals should be randomly scattered around zero, displaying no discernible pattern, and the variance of the residuals should remain consistent across all values of the independent variable (Field et al., 2012). However, since the residuals display a visible pattern, we use the Durbin-Watson test for autocorrelation to statistically verify the independence of errors. Typically, a statistic value less than one or greater than three is cause for concern (Field et al., 2012). Our plots and the Durbin-Watson test reveal a significant correlation between the observations of the error term. To address the violation of errors' independence, we fit the regression models with robust standard errors (HAC3). This standard statistical technique provides robust estimates of standard errors in the presence of autocorrelation (Cribari-Neto et al., 2007). Regarding homoscedasticity, the residuals seems to be evenly dispersed around the horizontal line. To corroborate this observation, we use the Breusch-Pagan Test (Breusch and Pagan, 1979), which confirms the absence of heteroscedasticity. A report of these test results is included in Appendix O.

Finally, we assess the normality of the residuals using QQ plots, the Shapiro-Wilk test, and the Kolmogorov-Smirnov (KS) test. Both analyses indicate a violation of the normality assumption in each regression model. However, the Central Limit Theorem (CLT) suggests that with a sufficiently large sample size (> 30), the distribution of the sample mean of random variables will approximate a normal distribution, regardless of the shape of the population distribution (Kwak and Kim, 2017). Since these models include 151 observations, the impact of non-normality is mitigated, allowing us to conduct a reliable statistical analysis despite the violation of normality. A report of these test results is included in Appendix O.

### 5.1.2 Results: Simple Regression Analysis

Adapting the approach of Hong et al. (2021), we first investigate the relationship between attitudes towards creative AI and the evaluation of songs with AI-cloned vocals using Models 1 and 2. Tables 17 report the output of the regression for both models. We find a significant and positive coefficient for both pop songs ($\beta = 0.224$, $p = 0.025$) and rock songs ($\beta = 0.238$, $p = 0.012$) at the 95% confidence level, suggesting that, when considered alone, there is evidence supporting a positive relationship. Specifically, if attitudes towards creative AI increase by one unit, we would expect the evaluation of pop songs to increase by 0.224 and that of rock songs by 0.238. This aligns with Hong et al. (2021), who identified a simple linear relationship between attitudes towards creative AI and the evaluation of its music. While their study focused on instrumental music in the EDM and classical genres, our findings extend this relationship to vocal music in the pop and rock genres. The magnitude of the relationship (effect size) is similar across both genres, consistent with Hong et al. (2021). However, the effect sizes are notably low, suggesting that other variables might confound this relationship. This is further supported by the low adjusted R-squared values, which imply that the models explain only a small portion of the variance in the dependent variable. As a result, before validating these conclusions we proceed with multiple regression analysis. Nevertheless, since we do not have sufficient evidence to reject H1 at this stage, *Attitudes_creative_AI* is included as a covariate in the multiple regression models.

Next, following the methodology of Hong (2022), we investigate the relationship between accepting AI in the role of a singer and the evaluation of songs with AI-cloned vocals using Models 3 and 4. Tables 17 report the output of the regression for both models. We find a significant and positive coefficient for both pop songs ($\beta = 0.288$, $p = 0.048$) and rock songs ($\beta = 0.253$, $p = 0.046$) at the 95% confidence level, implying that evidence supports a positive relationship when considered alone. Specifically, for every one-unit increase in acceptance of AI as a singer, we would expect the evaluation of pop songs to increase by 0.288 and that of rock songs by 0.253. These findings align with Hong (2022), who observed that when an AI is accepted as a musician, its musical compositions are appreciated more. Additionally, this may suggest that, when evaluating music, individuals hold a similar regard for the role of AI as both composer and singer. However, the effect sizes and adjusted R-squared values are notably low, indicating a weak relationship and limited explanatory power of the models. Consistent with Hong (2022), before drawing final conclusions about this relationship, we proceed with multiple regression analysis. Nonetheless, since we cannot reject H2 at this stage either, *Acceptance_AI_singer* is included as a covariate in our multiple regression models for further testing.

*Table 17: Simple Linear Regression Models Output*

| **Variable** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
|---|---|---|---|---|
| Attitudes_creative_AI | 0.224** | 0.238** | | |
| | (0.099) | (0.094) | | |
| Acceptance_AI_as_singer | | | 0.288** | 0.253** |
| | | | (0.145) | (0.126) |
| Intercept | 2.805*** | 2.927*** | 3.048*** | 3.293*** |
| | (0.380) | (0.368) | (0.330) | (0.298) |
| Observations | 151 | 151 | 151 | 151 |
| $R^2$ | 0.031 | 0.039 | 0.026 | 0.023 |
| Adjusted $R^2$ | 0.025 | 0.032 | 0.020 | 0.016 |
| F Statistic | 4.79** | 6.08** | 4.03** | 3.43** |

Notes: a) Robust Std. Errors are in parenthesis. HC3 Method.
b) ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

### 5.1.3 Robustness Check: Bootstrapping

To ensure the robustness of our findings, we perform bootstrapping on our four regression models. Bootstrapping is a non-parametric approach to statistical inference that does not require distributional assumptions, allowing us to assess the stability of our estimates despite violating normality (Fox, 2015). Following Field et al. (2012), we create 2000 bootstrap samples and report the 95% confidence intervals of the bootstrap coefficient estimates. Overall, bootstrapping our models confirms the robustness of our findings, with the bootstrapped confidence intervals reinforcing the significance and direction of our observed relationship. We report the output of regressions with bootstrapping in Appendix P.

### 5.1.4 Robustness Check: Spearman's Rank Correlation

Given the violation of the normality assumption, we also use Spearman's correlation coefficients as an additional robustness check for our regression results. Spearman's correlation is a non-parametric statistic measuring the monotonicity of relationships between variables by ranking the data and then applying Pearson's equation (Field et al., 2012). Overall, Spearman's rank coefficients are consistent with our regression results, reinforcing the observed relationships. We report the correlation matrix in Appendix P.

## 5.2 Hypotheses Testing

We test the causal effect of expectations, vocal quality, and their interaction on the evaluation of songs featuring AI-cloned vocals using multiple regression analysis. We first estimate two main effects models, one for each music genre, and then we include the interaction term between expectation and vocal quality (Table 18). Multiple regression extends simple linear regression and predicts the value of a dependent variable based on two or more independent variables. In this study, the dependent variable, *song_evaluation*, is treated as continuous despite its ordinal nature. This approach is supported by previous research that treats Likert scales as continuous for simpler interpretation (Harpe, 2015). We then include our two manipulated variables, coded as binary (1 or 0). The variable *Positive_expectations* represents expectations about the quality of AI-cloned vocals, with 1 indicating positive expectations and 0 indicating

negative expectations. The variable *High_quality* represents the quality of the vocals, coded as 1 for high-quality song samples and 0 for low-quality song samples. Additionally, informed by the analysis in Chapter 5.1, we include two covariates: *Attitudes_creative_AI*, and *Acceptance_AI_singer*. Although theoretical considerations suggested the inclusion of additional control variables (e.g., music genre preferences), findings from the preliminary analysis and assumption testing led to their exclusion. This is elaborated upon in Chapter 5.2.1.

*Table 18: Multiple Regression Models*

| Model | Regression Equation |
|---|---|
| Model 5 - Pop | $Song\_Evaluation = b_0 + b_1 \times expectations + b_2 \times vocal\_quality + b_3 \times attitudes\_creative\_AI + b_4 \times acceptance\_AI\_as\_singer + \varepsilon_i$ |
| Model 6 - Rock | $Song\_Evaluation = b_0 + b_1 \times expectations + b_2 \times vocal\_quality + b_3 \times attitudes\_creative\_AI + b_4 \times acceptance\_AI\_as\_singer + \varepsilon_i$ |
| Model 7 - Pop | $Song\_Evaluation = b_0 + b_1 \times expectations + b_2 \times vocal\_quality + b_3 \times attitudes\_creative\_AI + b_4 \times acceptance\_AI\_as\_singer + b_5 \times expectations*vocal\_quality + \varepsilon_i$ |
| Model 8 - Rock | $Song\_Evaluation = b_0 + b_1 \times expectations + b_2 \times vocal\_quality + b_3 \times attitudes\_creative\_AI + b_4 \times acceptance\_AI\_as\_singer + b_5 \times expectations*vocal\_quality + \varepsilon_i$ |

### 5.2.1 Assumptions: Multiple Linear Regression

Multiple linear regression relies on five main assumptions: a linear relationship between the independent and dependent variables, independent errors, homoscedasticity, no perfect multi-collinearity, and normally distributed errors (Field et al., 2012). To validate the use of this statistical technique and ensure the reliability of our findings, we test the assumptions for all regression models. Appendix R includes a report of these test results.

The first assumption requires the independent variables to be linearly related to the dependent variable. The analysis in Chapter 5.1 confirmed this is the case for *attitudes towards creative AI* and *acceptance of AI as a singer*. Our theoretical framework suggested that additional factors, such as music genre preferences, listening habits, and formal musical training, could influence our dependent variable and should be controlled for. We used scatter plots with

Loess lines and correlation coefficients to assess their relationship with our dependent variables. These analyses did not show a linear relationship; thus, these variables were excluded from further analysis. However, it is important to note that our randomisation procedure created comparable groups across these variables (Chapter 4.4); hence, their exclusion should not significantly impact the results.

Then, as with simple linear regression, we plotted the residuals against the fitted values of the regression models to visually inspect whether the residuals are randomly scattered around zero, display no discernible pattern, and maintain consistent variance across all values of the independent variable. While this appeared to be the case, we also used the Durbin-Watson and Breusch-Pagan tests to statistically corroborate our visual inference, checking for autocorrelation and heteroscedasticity, respectively. The tests reported no violation of these two assumptions, thus confirming the data is appropriate for multiple regression analysis.

To identify potential instances of multi-collinearity, we inspect the Variance Inflation Factor (VIF). The VIF estimates the degree to which a predictor is linearly associated with the other predictors (Field et al., 2012). Although there are no hard rules about what values of VIF should be problematic, Myers (1990) suggests that values exceeding 10 should raise concerns. In contrast, a value of 1 signifies no multi-collinearity (James et al., 2013). In our analysis, VIF values range from 1 to 1.46, and the reciprocal tolerance values (1/VIF) also fall within acceptable ranges.

Lastly, we need to verify whether the residuals display a normal distribution. The inspection of QQ plots and the results of the Shapiro-Wilk test for normality reporting insignificant coefficients indicate that the residuals approximate a normal distribution. Therefore, applying multiple regression analysis is valid and appropriate.

### 5.2.2 Results: Multiple Regression Analysis

A multiple regression analysis was conducted to statistically determine the effect of positive expectations and high-quality vocals - comparing them against negative expectations and low-quality vocals - on the evaluation of songs featuring AI-cloned vocals. We controlled for attitudes towards creative AI and acceptance of AI as a singer. The results of the regression

models are presented in Table 19. The overall F-test indicates that these models provide a better fit to the data than a baseline model that uses the mean of the dependent variable.

When examining *Attitudes_creative_AI*, the analysis reveals that there is not enough statistical evidence to conclude that changes in attitudes towards creative AI are associated with changes in the evaluation of songs with AI-cloned vocals. Specifically, the observed effect is not significantly different from zero in either pop ($\beta = 0.094$, $p = 0.192$) or rock ($\beta = 0.145$, $p = 0.162$) music. Although the simple linear regressions suggested a positive correlation to exist, when including the manipulated predictors and the other covariate, this relationship ceased to be relevant, thus indicating that different factors predict changes in song evaluation. These findings also contradict a previous study suggesting a positive relationship between this variable and the evaluation of AI instrumental music (Hong et al., 2021). Therefore, the specific case of music with AI vocals likely implies different dynamics than AI instrumental music, with other factors playing a more significant role in its evaluation. Consequently, we reject H1.

Similarly, for *Acceptance_AI_singer*, the results indicate that there is not enough statistical evidence to conclude that changes in the acceptance of AI as a singer are associated with changes in the evaluation of songs with AI-cloned vocals. Specifically, the observed effect is not significantly different from zero in either pop ($\beta = 0.126$, $p = 0.209$) or rock ($\beta = 0.063$, $p = 0.577$) music. Similar to *Attitudes_creative_AI*, the simple linear regressions suggested a positive correlation to be present. However, including the manipulated variables and the other covariates revealed that other factors play a more significant role in predicting song evaluation. These findings also contradict Hong (2022), where higher acceptance of an AI composer was found to positively influence the assessment of its music. This may imply that the evaluation of music with AI vocals involves different dynamics than AI instrumental music and that the role of a singer is regarded differently than that of a composer. Consequently, we reject H2.

We then assess the effect of our manipulated variables, *expectations* and *vocal quality*, on song evaluations. The results indicate that having positive expectations, compared to negative expectations, does not influence the evaluation of songs with AI-cloned vocals. Specifically, the effect is not statistically different from zero in either pop ($\beta = 0.224$, $p = 0.457$) or rock ($\beta = 0.131$, $p = 0.728$) music. While previous studies suggested that having positive ex-

pectations influences the evaluation of AI instrumental music (Hong et al., 2021) and intention to adopt AI products (Hong, 2021), our data does not lead to similar conclusions across music with AI vocals. The discrepancy in findings may arise since these studies were observational, whereas our results arise from an experimental setting. Consequently, we reject H3.

Lastly, when inspecting *High_quality*, we find enough statistical evidence to conclude that high-quality AI vocals, compared to low-quality AI vocals, significantly influence the evaluation of songs in both music genres at the 99% confidence level. Specifically, when songs are of high quality, compared to low quality, they are evaluated, on average, 2.762 units higher in the case of pop music ($\beta = 2.762$, $p < 0.01$), ceteris paribus, and 2.423 units higher in the case of rock music ($\beta = 2.423$, $p < 0.01$), ceteris paribus. These results are in line with previous studies indicating that the quality of AI products, including AI-generated artworks and music, is a leading factor in their evaluation, regardless of their origin - i.e., whether they are human or AI-generated (Hong and Curran, 2019; Shank et al., 2023). Consequently, since moving from low-quality to high-quality vocals results in significantly higher song evaluations regardless of the expectation level, we find support for H4.

Our experimental design also allows us to test our fifth hypothesis indirectly. Specifically, we have created four groups based on two levels of expectations and two levels of vocal quality. The different levels of vocal quality ensure that Group 1 and Group 3 differ in terms of expectancy violation, as do Group 2 and Group 4, as confirmed in Chapter 4.5. By including an interaction term in our regression model, we test whether the combined effect of expectations and vocal quality significantly influences the dependent variable. In other words, we are examining whether the impact of expectations on song evaluation changes across different levels of vocal quality. Since we confirmed that different expectancy violations occur when the expected and vocal quality differ, the interaction term indirectly captures the effect of this violation. However, when inspecting the output of Models 7 and 8 (Table 19) we do not find enough statistical evidence to infer that the effect of the interaction term is significantly different from zero in either pop ($\beta = -0.032$, $p = 0.323$) and rock ($\beta = 0.085$, $p = 0.353$) music. These findings find partial support in the literature. Indeed, expectancy violations were found to have different effects on AI instrumental music depending on whether the genre was classical

or EDM. Our study finds that EVT applies to pop and rock vocal music similarly to EDM, with no significant changes in evaluations based on different levels of violations. Consequently, we reject H5. Regarding the effects of the other independent variables and covariates in Models 7 and 8, we find consistent results to those in Models 5 and 6, with parallel interpretations.

Overall, these findings suggest that the quality of AI-cloned vocals plays a crucial role in how songs with AI-cloned vocals are evaluated, overshadowing the effects of initial expectations, attitudes towards AI in the creative field and acceptance of AI in the singer's role. However, we acknowledge that we are only indirectly testing for the moderation effect of expectancy violation, and the limitations of this approach are further discussed in Chapter 7. Nonetheless, we also provide a robustness check by directly testing for the moderation effect using the values of the variable *expectancy_violation* collected in our survey, employing Hayes' PROCESS macro for SPSS (Chapter 5.2.4).

*Table 19: Multiple Linear Regression Models Output*

| **Variable** | **Model 5** | **Model 6** | **Model 7** | **Model 8** |
|---|---|---|---|---|
| Positive_expectations | 0.224 | 0.131 | 0.240 | 0.088 |
| | (0.162) | (0.175) | (0.231) | (0.250) |
| High_quality | 2.762*** | 2.423*** | 2.777*** | 2.390*** |
| | (0.162) | (0.175) | (0.224) | (0.243) |
| Positive_Expectations*High_quality | | | -0.032 | 0.085 |
| | | | (0.326) | (0.352) |
| Attitudes_creative_AI | 0.094 | 0.145 | 0.094 | 0.146 |
| | (0.071) | (0.077) | (0.071) | (0.077) |
| Acceptance_AI_singer | 0.126 | 0.063 | 0.127 | 0.061 |
| | (0.100) | (0.108) | (0.100) | (0.109) |
| Intercept | 1.530*** | 1.859*** | 1.523*** | 1.878*** |
| | (0.251) | (0.271) | (0.263) | (0.285) |
| Observations | 151 | 151 | 151 | 151 |
| $R^2$ | 0.828 | 0.769 | 0.828 | 0.769 |
| Adjusted $R^2$ | 0.685 | 0.685 | 0.580 | 0.592 |
| F Statistic | 79.5*** | 52.9 *** | 63.2*** | 42.0*** |

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

### 5.2.3 Robustness Check: ANCOVA

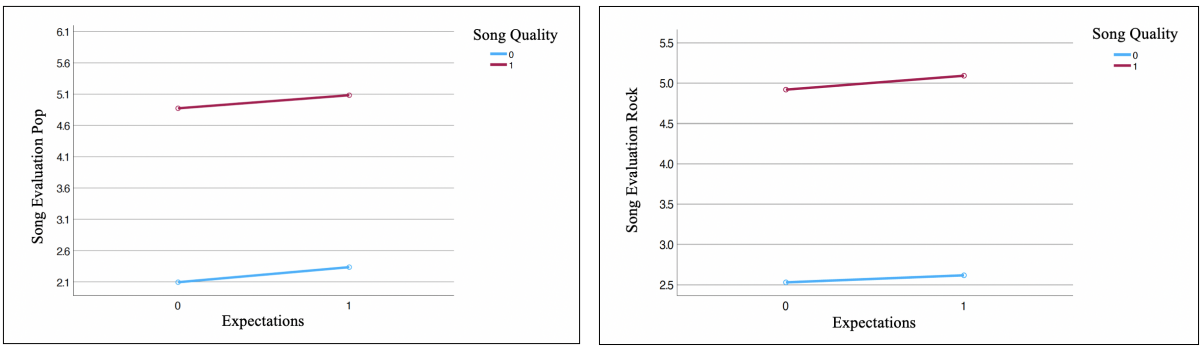
We validate the results of Chapter 5.2.2 using two sets of two-way analysis of covariance (ANCOVA), one for each genre condition. ANCOVA extends ANOVA (analysis of variance) to include covariates. While regressions are typically used to determine the strength of the relationship between variables, making them particularly suited for testing H1 and H2, ANCOVA is typically used in experimental research as a special case of regression analysis to compare means of categorical independent variables and a continuous dependent variable (Field et al., 2012), making it particularly suited for testing H3, H4, and H5. Therefore, we use ANCOVA as a robustness check.

#### 5.2.3.1 Assumptions: ANCOVA

ANCOVA requires the data to comply with the same assumptions as regression analysis, along with two additional assumptions: homogeneity of regression slopes and homogeneity of variances. Ensuring these assumptions are met is crucial for the validity and reliability of the ANCOVA results. The first additional assumption, homogeneity of regression slopes, implies that the relationship between each covariate and the dependent variable is consistent across all groups (Field et al., 2012). We inspected scatter plots and found no reason to suspect a violation of this assumption. Moreover, we statistically confirmed this performing a test of homogeneity of regression slopes in SPSS (Appendix S). The second, homogeneity of variances, means that the variance within each group is similar across all groups (Field et al., 2012). We performed Levene's test of equality of variances and found no significant violations ($p = 0.517$ and $p = 0.433$). Besides testing all other assumptions again following the same methodology used in Chapter 5.2.2, we additionally verify that they hold within each experimental condition. The only relevant result from this latter approach comes from the Shapiro-Wilk test, which indicates non-normally distributed errors within one experimental condition (as we suspected based on the visual inspection in Chapter 4.6). However, because the overall model residuals are normally distributed and we can rely on the CLT ($N > 30$) for non-normality within the condition, we can proceed with ANCOVA. A report of these test results is included in Appendix S.

#### 5.2.3.2 Results: ANCOVA

The results from the ANCOVA corroborate the findings of Chapter 5.2.2. We first examine the interaction effect by creating interaction plots, which demonstrates no interaction in either pop (Figure 7a) or rock (Figure 7b) music. Examining the interaction effect first is crucial because, if significant, it invalidates the main effects, as the factors would be working together to affect the dependent variable (Field et al., 2012). The ANCOVA output tables (Tables 20 for pop and 21 for rock) confirm our inference, showing no significant interaction effect in either pop ($F[0.01]$, $p = 0.923$) or rock ($F[0.06]$, $p = 0.811$) music. Next, when assessing the main effects, we again find no significant effect for expectations on song evaluation in either pop ($F[0.55]$, $p = 0.169$) or rock ($F[1.92]$, $p = 0.461$) music. Consistent with the regression results, we find a significant effect of vocal quality in both the pop ($F[288.57]$, $p < .001$, $\eta^2$ = 0.67 ) and rock ($F[190.89]$, $p < .001$, $\eta^2$ = 0.57) genre, at the 99% confidence level. Specifically, the adjusted means reported in Appendix S show that the high-quality vocal group has significantly higher evaluations than the low-quality vocal group for both pop ($M = 4.998$ vs. $M = 2.237$) and rock music ($M = 5.006$ vs. $M = 2.574$). As for the covariates, we again find no significant effects ($p > 0.1$) (Tables 20 and 21), indicating they do not adjust the association between the factors and the dependent variable. In light of these results, ANCOVA corroborates the findings of the multiple regression analysis, leading us to reject H1, H2, H3, and H5 and to support H4.




*(a) Pop* *(b) Rock*

*Figure 7: ANCOVA interaction plots*

*Table 20: ANCOVA Output Pop*

| **Source** | **SS** | **df** | **MS** | **F** | $\eta^2$ |
|---|---|---|---|---|---|
| Model | 2288.32 | 6 | 381.39 | 389.43*** | 0.94 |
| Attitudes_creative_AI | 1.68 | 1 | 1.68 | 1.72 | 0.01 |
| Acceptance_AI_singer | 1.56 | 1 | 1.56 | 1.59 | 0.01 |
| Positive_expectations | 1.88 | 1 | 1.88 | 1.92 | 0.01 |
| High_quality | 282.61 | 1 | 282.61 | 288.57*** | 0.67 |
| Positive_expectations * High_quality | 0.01 | 1 | 0.01 | 0.01 | 0.00 |
| Residual | 142.00 | 145 | 0.98 | | |
| Total | 2430.33 | 151 | | | |

Notes: a) ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
b) $R^2$ = 0.942, Adjusted $R^2$ = 0.939
c) Type III Sum of Squares (SS)

*Table 21: ANCOVA Output Rock*

| **Source** | **SS** | **df** | **MS** | **F** | $\eta^2$ |
|---|---|---|---|---|---|
| Model | 2418.34 | 6 | 403.06 | 350.81*** | 0.94 |
| Mean_score_attitudes_creative | 4.064 | 1 | 4.064 | 3.54 | 0.02 |
| Mean_score_acceptance_role | 0.36 | 1 | 0.36 | 0.31 | 0.00 |
| Positive_expectations | 0.63 | 1 | 0.63 | 0.55 | 0.00 |
| Good_song | 219.32 | 1 | 219.32 | 190.89*** | 0.57 |
| Positive_expectations * Good_song | 0.07 | 1 | 0.07 | 0.06 | 0.00 |
| Error | 166.60 | 145 | 1.15 | | |
| Total | 2584.94 | 151 | | | |

Notes: a) ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
b) $R^2$ = 0.936, Adjusted $R^2$ = 0.933
c) Type III Sum of Squares (SS)

### 5.2.4 Robustness Check: Hayes' PROCESS Macro

The PROCESS macro by Andrew F. Hayes is a tool designed for conducting various forms of regression analysis, particularly useful for complex models such as moderation and mediation analysis (Hayes and Rockwood, 2017). It automatically calculates interaction effects at different levels of the moderator, simplifying the process, reducing the risk of errors, and providing detailed output, which offers a deeper understanding of the relationships between variables (Regorz, 2021). Since our previous analysis only indirectly addressed H5, we now use the values of *expectancy_violation* to directly assess the potential moderating effects of stronger expectancy violations. Moreover, since we manipulated and identified two types of expectancy violations (refer to Chapter 3.3.6 and 4.5), one statistically stronger in a positive direction between groups 2 and 4, and the other statistically stronger in a negative direction between groups 1 and 3, we split the data accordingly and performed separate analyses.

First, we examined the interaction effect and found that it is not statistically significant ($p > 0.1$) in any of our analyses. Then, we assessed the conditional effect of expectations on song evaluations at three levels of *expectancy_violation* ($-1SD$, Mean, $+1SD$), and again, observed insignificant effects ($p > 0.1$). As an additional check, we plotted the conditional effect of expectations, which confirmed our statistical findings, with lines not intersecting. Lastly, we repeated the same analysis with 2000 bootstrap samples to address potential bias arising from the slightly unbalanced group sample sizes (refer to Chapter 4.7). Hayes and Rockwood (2017) also recommend the use of bootstrap samples in small sample sizes. Based on this analysis, we draw parallel conclusions. Consequently, consistent with the regression analysis in Chapter 5.2.2, expectancy violation does not significantly moderate the relationship in either pop or rock music at any of its levels, in either direction, leading us to reject H5 again. For reference, we report the specific SPSS settings used to run the analyses in Appendix T.

# 6 Discussion

This chapter addresses our research questions, discusses the results of the analysis and compares them against previous studies. Additionally, the chapter delves into the contributions of this research, highlighting both academic and managerial implications.

## 6.1 Findings and Implications

This study aimed to explore public reactions to songs with AI-cloned vocals by addressing how individuals evaluate them under varying expectations and levels of vocal quality. The study also investigated the effect of inherent attitudes towards AI in creative endeavours and beliefs about AI cloning the voices of singers in shaping these reactions. Additionally, we tested two popular and contrasting music genres to understand whether the results apply broadly across musical types. With these objectives in mind, our study was centred around the following research question:

**RQ1**: How do expectations about songs featuring AI-cloned vocals influence people's evaluation of such songs?

Furthermore, our experimental design enabled us to investigate an additional element typically relevant when addressing expectations. Specifically, we examined whether stronger expectancy violations influenced the evaluation of songs by moderating the effect of expectations. Accordingly, central to our study was the following research question:

**RQ2**: How does violating expectations about songs with AI-cloned vocals influence people's evaluation of such songs?

Drawing on relevant literature, we formulated and tested five hypotheses to address our research questions. Table 22 provides an overview of our hypotheses and the analysis outcomes. Overall, we find support for H4, whereas we do not find enough statistical evidence to support H1, H2, H3, and H5.

*Table 22: Hypotheses Overview*

| Hypothesis | Outcome |
|---|---|
| **H1:** The more positive the attitudes towards creative AI, the higher the evaluations of songs with AI-cloned vocals. | **Rejected** |
| **H2:** The higher the acceptance of AI as a singer, the higher the evaluations of songs with AI-cloned vocals. | **Rejected** |
| **H3:** Having positive expectations about the quality of AI-cloned vocals will lead to higher evaluations of songs with AI-cloned vocals compared to having negative expectations. | **Rejected** |
| **H4:** High-quality vocals will lead to higher evaluations of songs with AI-cloned vocals than low-quality vocals. | **Supported** |
| **H5:** The effect of expectations on the evaluation of songs with AI-cloned vocals is stronger with greater levels of expectancy violations. | **Rejected** |

Our experimental findings suggest that expectations do not influence people's evaluation of songs with AI-cloned vocals; instead, it is the quality of the vocals that ultimately matters. These results offer partial support to the existing literature. While positive expectations have been shown to improve the evaluation of AI instrumental music (Hong et al., 2021) and foster the adoption of AI products (Hong, 2021), our data indicates otherwise. The evaluations of AI vocal music remain unaffected by initial expectations, suggesting that it operates under different dynamics than AI instrumental music.

The quality of the AI vocals appears to be a decisive factor in this context. Indeed, transitioning from low-quality to high-quality vocals results in significantly higher evaluations. In other words, people evaluate songs positively if the AI-cloned vocals are of high quality and negatively if they are of low quality. This aligns with previous studies indicating that the quality of AI products, including AI-generated artworks (Hong and Curran, 2019) and music (Shank et al., 2023), is a leading determinant of their evaluation, regardless of their origin - whether human or AI-generated. Quality is a primary measure of value and satisfaction in consumer

products. However, this association can be less consistent with AI products or human-computer interactions (Hong, 2022). Thus, validating this factor's importance in this context substantially contributes to this technology's development and practical applications.

Our analysis extended to other potential influences, but the data did not support our hypotheses. Specifically, we observed that neither believing in the creative abilities of AI nor accepting AI as a legitimate singer is associated with the evaluation of its music, whether of high or low quality. This contrasts with Hong et al. (2021), who reported a positive correlation between attitudes towards creative AI and the evaluation of its music, and Hong (2022), where greater acceptance of AI composers positively influenced their music evaluations. This discrepancy reinforces our proposition that the evaluation of AI vocal music is subject to different dynamics than AI instrumental music, with the roles of singers and composers being viewed differently. It is speculative, but AI musicians and AI singers may be seen differently in terms of the human-like traits required for each role. In other words, anthropomorphism - attributing human characteristics to non-human entities - could influence how people accept AI in these roles. While people may be more open to AI creating instrumental music, a role viewed as more technical, they may be less accepting of AI as singers, a role closely associated with human emotions. Indeed, there appears to be an underlying sentiment questioning the artistic merit of AI-generated vocals. Participants generally reported low acceptance of AI taking over the role of singers. Although this sentiment does not affect song evaluations, it suggests that people may view singing as a uniquely human trait that machines should not replicate.

We also found no evidence supporting a moderating role of expectancy violations. Being surprised by the quality of the AI vocals, whether much better or much worse than anticipated, does not influence evaluations. These findings add complexity to the Expectancy Violation Theory, partially aligning with prior studies focused on instrumental music. Specifically, expectancy violations were found to have varied effects on AI instrumental music depending on whether the genre was classical or EDM (Hong et al., 2021). Our study indicates that EVT functions similarly in pop and rock vocal music as it does in EDM, with no notable changes in evaluations based on different levels of violations. This was unexpected because pop and rock music might be seen as more akin to classical music, which is often perceived as more human-

centred, whereas EDM is seen as more synthetic. Again, the introduction of AI vocals appears to present novel dynamics in how listeners evaluate these songs, diverging from the patterns observed with AI instrumental music, which future research should address and uncover.

Although rock and pop music typically elicit different emotional responses, with rock enthusiasts being associated with more openness to innovation than pop fans (Rentfrow and Gosling, 2003), our findings extend equally across both genres. This further reinforces that vocal quality significantly outweighs other elements when evaluating songs with AI vocals. Nevertheless, it is possible that the experimental design may have contributed to such findings by polarising evaluations based on quality, leading to an insignificant effect of other elements. We further address this aspect in Chapter 7.

In the broader discussion of what constitutes creativity (Coeckelbergh, 2017), our findings support that songs with AI-cloned vocals are creative works if art is defined using objective criteria. Indeed, their evaluations are positive when the AI is engineered to produce indistinguishable vocals. However, when the definition of art relies on subjective criteria, it is still uncertain whether songs with AI-cloned vocals would qualify as art or, at least, accepted on the same level as human songs.

### 6.1.1 Academic Implications

In addition to what is already addressed in Chapter 6.1, this study significantly contributes to the existing literature on AI applications within the creative domain. It is the first empirical study to explore public perceptions of songs featuring AI-cloned vocals, shedding light on critical elements that could impact technology acceptance. Additionally, it adds to the broader discussion of human-computer interactions (HCI) within the musical domain. It also adds to the body of literature moving towards understanding AI's impact on music from a social and psychological stance.

Our findings, though not entirely consistent with some prior studies, benefit from an experimental approach. Most studies on creative AI and EVT have been observational. In contrast, our methodology allows us to infer causality with greater confidence, which may also explain the differences in results. Nevertheless, the lack of support for the Expectancy

Violation Theory (EVT) warrants further discussion. While this study enriches EVT by offering insights into how it applies to this specific context, additional research should address the 'why' - specifically, why EVT operates differently depending on whether AI assumes the role of composer or singer.

### 6.1.2 Managerial Implications

Our findings reveal that the evaluation of songs featuring AI-cloned vocals is primarily determined by their technical quality. This insight offers important implications for music producers, technology developers, and marketers. First, since initial expectations do not significantly impact song evaluations, marketing strategies should center on showcasing the AI's ability to produce high-quality, realistic vocals to secure public acceptance, rather than attempting to shape expectations. Second, the critical role of vocal quality suggests that investments in enhancing the technical quality of AI-generated vocals should be prioritised. High-quality vocals are more likely to receive positive evaluations, highlighting the need for continuous improvements in AI voice synthesis technology. Third, the low acceptance of AI as legitimate singers indicates a need to address concerns about the artistic merit of AI-generated vocals. Engaging in conversations with the public and integrating their feedback into the development process can help align technological progress with public perception. Lastly, a positive evaluation may not directly correlate with a willingness to listen to or purchase music featuring AI-cloned vocals. While one might positively evaluate a song that clones the voice of Taylor Swift when asked to report an assessment, it remains uncertain whether they would actively choose to listen to such a product regularly or even replace their listening habits entirely. Our findings suggest that there are still many unaddressed elements that might influence the acceptance of this technology, which further research should address before making broader conclusions. Nonetheless, at this stage, it is evident that much progress and research are needed before AI-generated vocals can rival established artists like Taylor Swift.

# 7 Limitations & Future Research

Drawing on relevant literature, this study has been carefully designed to produce valid and reliable results. Nonetheless, it presents several inherent limitations. This chapter seeks to address these limitations while providing suggestions for future research. Specifically, we discuss limitations related to the study's internal and external validity and the econometric models used. We offer suggestions to address these limitations and outline future research opportunities at the intersection of AI and music.

## 7.1 Validity & Generalisability

Data collected via Prolific is comparable to that from laboratory studies (Palan and Schitter, 2018; Peer et al., 2017) and is superior to rival platforms such as Amazon Mechanical Turk (Peer et al., 2022). However, using crowdsourcing platforms might still produce low-quality responses. Specifically, on Prolific, participants often engage in multiple studies per day, which could increase their tendency to speed through tasks, provide straight-line responses, and display decreased attention. Prolific recommends a compensation rate between £9.00 and £12.00 per hour to foster higher engagement and ensure high data quality (Denison, 2023), whereas we were only able to offer £6.00/hr. Nonetheless, we included survey elements designed to detect low-quality responses and address survey-related biases (refer to Chapters 3.3.8 and 3.3.9).

Using Prolific might also pose limits on the study's generalisability. Indeed, Prolific predominantly uses convenience sampling, where studies are filled on a first-come, first-served basis (Prolific, 2023). As a result, launching our survey at different times of the day or on different days of the week could have produced a sample population with other characteristics. Moreover, participant pools on Prolific are skewed towards Western, Educated, Industrialised, Rich, and Democratic (WEIRD) individuals, who are typically more technologically savvy and familiar with AI (Henrich et al., 2010). We observed a similar trend in our sample population (refer to Chapter 4.1), with the over-representation and under-representation of specific population groups. Consequently, the study's ability to generalise to other populations is limited.

Then, while we tried to standardise the listening experience as much as possible by recommending the use of headphones, we were unable to control hardware use. If participants listened to the song samples using different devices, their listening experiences may have differed, potentially affecting the results. Future research should replicate these findings in a laboratory setting to have more control over the use of standardised equipment.

Furthermore, the song samples used were 15s long. Although Belfi et al. (2018) show this should be sufficient to make stable evaluations, it may not fully represent how people listen to music in real-world settings. Additionally, pop and rock music encompass dozens of sub-genres, each with distinct characteristics. Since we only provided two song samples for each genre, this study can only make inferences about a specific pop and rock music subcategory. Moreover, pop and rock are part of the Western musical tradition. Historically, music research has often ignored non-Western music (Baker et al., 2020). Hence, future research could replicate these findings by observing song evaluations in more naturalistic settings and over other genres or sub-genres.

While the study addressed the evaluation of songs with AI-cloned vocals, we used actual human vocals but informed participants they were AI-generated. Using actual AI-generated vocals would provide more accurate and applicable insights. However, due to limited access to high-quality AI-cloned vocals, this approach simplified the experimental procedure while still providing valid results, as the technology has already shown instances of producing realistic and indistinguishable vocals. Moreover, this method allowed us to directly manipulate the vocals to sound robotic and create two distinct conditions with greater precision.

To prime expectations, we provided fictional information about AI cloning technology in a scenario-based experiment (refer to Chapter 3.3.3). While numerous studies have adopted this approach, providing reliable results (Gregory and Duran, 2001; Olshavsky and Miller, 1972), priming expectations in controlled, static settings may not accurately reflect how people form expectations in real-world scenarios. Besides, if participants already hold strong pre-existing expectations or have had experiences with AI-cloned vocals, the priming information might have a limited impact. Nonetheless, we confirmed that the manipulation was, at least, effective in differentiating the groups.

Drawing on previous literature, we employed a within-subject design to examine both pop and rock music to increase generalisability while keeping the study within our budgetary limitations. Despite using counterbalancing (Chapter 3.1.1), listening to one genre after the other may still have resulted in order, carryover, or learning effects (Jhangiani et al., 2019), thus biasing the evaluation of songs and limiting the causal validity of the findings. Future studies should consider including music genres as part of a between-subject design to provide a more reliable comparison of the effect of expectations or other manipulated factors on the evaluation of music.

In this context, directly manipulating expectancy violations may not be feasible, as one cannot control participants' cognitive processes in how they perceive and react to musical stimuli. Indeed, previous studies have primarily observed expectancy violations in correlational studies (Hong et al., 2021; Hong, 2022; Messingschlager and Appel, 2023; Waddell, 2018). Nevertheless, we chose to create contrasting conditions as a manipulative tool, successfully creating different groups in terms of expectancy violations. We acknowledge that this approach poses limitations on the causal inferences we can make about expectancy violations. Additional research is needed to validate our findings and further explore the causal relationships involved.

Moreover, while we found no significant effect of expectations on the evaluation of music with AI-cloned vocals, it is possible that we polarised the vocal quality to such an extent that it overshadowed the role of expectations. In other words, the low-quality vocals may have been too poor in quality, thereby dominating participants' evaluations. Future research might attempt to replicate these findings with varying levels of vocal quality to better understand the influence of expectations.

Lastly, being the first empirical study on AI-cloned vocals, our main objective was to assess general public reactions. However, we recognize that the technology's ability to clone a specific artist was not prioritised in this study. By intentionally selecting unfamiliar songs and artists, we aimed to gauge overall perceptions of the technology, though different reactions might arise when this technology is used to replicate beloved or popular artists, an area worthy of future exploration.

## 7.2 Econometric Models

Regarding the econometric models employed, we have addressed the violation of assumptions in Chapters 5.1.1, 5.2.1, and 5.2.3. While the use of bootstrapping in simple linear regression analysis and robust standard errors is valid, it may also contribute to overfitting and sampling error. The presence of non-normal residuals in the ANCOVA within one experimental condition may also be a source of concern, potentially affecting the validity of the results. The Central Limit Theorem allowed us to proceed with this statistical technique, as our groups each included between 30 and 40 participants. However, increasing the sample size and achieving even group sizes could lead to more accurate and reliable models. Furthermore, we treated our dependent variable as continuous despite its ordinal nature to simplify interpretation and facilitate the use of linear regression. We believe that providing meaningful results outweighs the potential biases arising from this approach.

As a final remark, most control variables did not display a linear relationship with the dependent variable and were thus excluded from the main models and only included as a robustness check in Appendix Q. This was unexpected and contradicted theoretical predictions; hence, future research may investigate it further. For instance, while we used the STOMP scale to capture music genre preferences (Rentfrow and Gosling, 2003), future studies could use the revised version, STOMP-R, which includes more genres and may capture genre preferences more precisely (Rentfrow and Gosling, 2011).

## 7.3 Future Directions

The field of AI and music remains largely unexplored, particularly concerning AI-cloned vocals and public perceptions. Indeed, the technology has only recently advanced to enter the public domain. Consequently, when reflecting on the present level of technological maturity, while experimental research is particularly useful for testing hypotheses and theories, future studies might adopt a more exploratory approach using qualitative methods such as interviews, focus groups, and case studies.

Moreover, our study has established that expectations do not play a significant role in people's evaluations; instead, vocal quality is the decisive factor. However, positive evaluations

and high-quality vocals do not directly imply that people will choose to listen to these songs. Accordingly, the recommended next step for industry practitioners is to investigate the extent to which people are willing to listen to or purchase such music and under which conditions or settings.

Lastly, beyond these considerations, the development of an appropriate regulatory framework to govern the technology is a pressing matter. The need for suitable laws, the review of copyright concepts, and the establishment of compensating systems are critical areas of research that must be addressed before the technology can be made widely available to the public.

# 8 References


Aïmeur, E., Amri, S. and Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(30). https://doi.org/10.1007/s13278-023-01028-5

Amezaga, N. and Hajek, J. (2022). Availability of Voice Deepfake Technology and its Impact for Good and Evil. *In Proceedings of the 23rd Annual Conference on Information Technology Education*, pp. 23-28. https://doi.org/10.1145/3537674.3554742

Anantrasirichai, N. and Bull, D. (2021). Artificial intelligence in the creative industries: a review. *Artificial Intelligence Review*, 55, pp. 589-656. https://doi.org/10.1007/s10462-021-10039-7

Arik, S., Chen, J., Peng, K., Ping, W. and Zhou, Y. (2018). Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems*, 31. https://doi.org/10.48550/arXiv.1802.06006

Baker, D. J., Belfi, A., Creel, S., Grahn, J., Hannon, E., Loui, P., Margulis, E. H., Schachner, A., Schutz, M., Shanahan, D. and Vuvan, D. T. (2020). Embracing anti-racist practices in the music perception and cognition community. *Music Perception*, 38(2), pp. 103-105. https://doi.org/10.1525/mp.2020.38.2.103

Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S. and Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, 18(2), pp.127-131. https://doi.org/10.4103/0972-6748.62274

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, 3, pp. 77-85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Belfi, A. M., Kasdan, A., Rowland, J., Vessel, E. A., Starr, G. G. and Poeppel, D. (2018). Rapid timing of musical aesthetic judgments. *Journal of Experimental Psychology: General*, 147(10), pp. 1531-1543. https://doi.org/10.1037/xge0000474

Bell, S. A. (2023). Federal Support for the Development of Speech Synthesis Technologies: A Case Study of the Kurzweil Reading Machine. *Information & Culture*, 58(1), pp. 39-65. https://doi.org/10.7560/IC58103

Benesty, J., Sondhi, M. M. and Huang, Y. A. (2008). Introduction to Speech Processing. In: Benesty, J., Sondhi, M. M. and Huang, Y. A. (eds.) *Springer Handbook of Speech Processing*, pp. 1-4. https://doi.org/10.1007/978-3-540-49127-9_1

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. 2nd edition. Tampa, Florida: University of South Florida.

Biddle, B. J. (1986). Recent developments in role theory. *Annual Review of Sociology*, 12(1), pp. 67-92. https://doi.org/10.1146/annurev.so.12.080186.000435

Birtchnell, T. and Elliott, A. (2018). Automating the black art: Creative places for artificial intelligence in audio mastering. *Geoforum*, 96, pp. 77-86. https://doi.org/10.1016/j.geoforum.2018.08.005

Blaauw, M. and Bonada, J. (2017). A neural parametric singing synthesizer modeling timbre and expression from natural songs. *Applied Sciences*, 7(12), p. 1313. https://doi.org/10.3390/app7121313

Bonito, J. A., Burgoon, J. K. and Bengtsson, B. (1999). The role of expectations in human-computer interaction. *In Proceedings of the 1999 ACM International Conference on Supporting Group Work*, pp. 229-238. https://doi.org/10.1145/320297.320324

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pp. 1287-1294. https://doi.org/10.2307/1911963

Burgoon, J. K. (1978). A communication model of personal space violation: Explication and an initial test. *Human Communication Research*, 4(2), pp. 129-142. https://doi.org/10.1111/j.1468-2958.1978.tb00603.x

Burgoon, J. K. (2015). Expectancy violations theory. *The International Encyclopedia of Interpersonal Communication*, pp. 1-9. https://doi.org/10.1002/9781118540190.wbeic102

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E. and Giboney, J. S. (2016). Application of Expectancy Violations Theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies*, 91, pp. 24-36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Casler, K., Bickel, L. and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), pp. 2156-2160. https://doi.org/10.1016/j.chb.2013.05.009

Cerence (2019). Cerence Introduces My Car, My Voice - New Voice Clone Solution to Personalize the In-Car Voice Assistant. *GlobeNewswire*. 30 December. Available at: https://bit.ly/MyCarMyVoice (Accessed: 2 June 2024).

Chamberlain, R., Mullin, C., Scheerlinck, B. and Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2), pp. 177-192. https://doi.org/10.1037/aca0000136

Charness, G., Gneezy, U. and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, 81(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Chen, X., Chu, W., Guo, J. and Xu, N. (2019). Singing voice conversion with non-parallel data. *In 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 292-296. https://doi.org/10.1109/MIPR.2019.00059

Chin, T. and Rickard, N. S. (2012). The music USE (MUSE) questionnaire: An instrument to measure engagement in music. *Music Perception*, 29(4), pp. 429-446. https://doi.org/10.1525/mp.2012.29.4.429

Cho, E. and Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2), pp. 207-230. https://doi.org/10.1177/1094428114555994

Choi, S., Kim, W., Park, S., Yong, S. and Nam, J. (2020). Korean singing voice synthesis based on auto-regressive boundary equilibrium gan. *In ICASSP International Conference on Acoustics, Speech and Signal Processing*, pp. 7234-7238. https://doi.org/10.1109/ICASSP40776.2020.9053950

Civita, 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*, 109. https://doi.org/10.1016/j.eswa.2022.118190

Coeckelbergh, M. (2017). Can machines create art? *Philosophy & Technology*, 30(3), pp. 285-303. https://doi.org/10.1007/s13347-016-0231-5

Cohen, L. and Reid, T. (2023). An early look at the possibilities as we experiment with AI and Music. *Youtube Official Blog*. 16 November. Available at: https://bit.ly/youblogAi (Accessed: 3 June 2024).

Coscarelli, J. (2023). An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World. *The New York Times*. 24 April. Available at: https://nyti.ms/3ultvom (Accessed: 10 February 2024).

Covach, J. (1997). We Won't Get Fooled Again: Rock Music and Musical Analysis. *In Theory Only*, 13(1-4), pp. 119-141. http://dx.doi.org/10.17613/wdm5-ka97

Cribari-Neto, F., Souza, T. C. and Vasconcellos, K. L. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics - Theory and Methods*, 36(10), pp. 1877-1888. https://doi.org/10.1080/03610920601126589

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), pp. 297-334. https://doi.org/10.1007/BF02310555

Davies, J., Klinger, J., Mateos-Garcia, J. and Stathoulopoulos, K. (2020). AI and the Creative Industries: The art in the artificial *Creative Industries Policy and Evidence Centre*. 11 June. Available at: https://pec.ac.uk/research-reports/the-art-in-the-artificial (Accessed: 12 June 2024)

Denison, G. (2023). How much should you pay research participants? 24 October. Available at: https://bit.ly/3SuTrH1 (Accessed: 12 June 2024).

Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A. and Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint*. https://doi.org/10.48550/arXiv.2005.00341

Donahue, B. (2024). Tupac Shakur's Estate Threatens to Sue Drake Over Diss Track Featuring AI-Generated Tupac Voice. *Billboard*. 24 April. Available at: https://bit.ly/3z9KZ8X (Accessed: 3 June 2024).

Drott, E. (2020). Copyright, compensation, and commons in the music AI industry. *Creative Industries Journal*, 14(2), pp.190-207. https://doi.org/10.1080/17510694.2020.1839702

Durbin, J. and Watson, G. S. (1992). Testing for serial correlation in least squares regression. II. *In Breakthroughs in Statistics: Methodology and Distribution*, pp. 237-259. https://doi.org/10.1007/978-1-4612-4380-9_20

Edwards, C., Edwards, A., Spence, P.R. and Westerman, D. (2016). Initial interaction expectations with robots: testing the human-to-human interaction script. *Communication Studies*, 67(2), pp. 227-238. https://doi.org/10.1080/10510974.2015.1121899

European Commission (2024). Shaping Europe's digital future | AI Act. *European Commission*. 19 June. Available at: https://bit.ly/4eQbQHu (Accessed: 20 June 2024).

European Commission (n.d.) Culture and Creativity | Music Moves Europe. *European Commission*. Available at: https://bit.ly/4d1Aljg (Accessed: 20 June 2024).

Eurostat (2024a). *Population by educational attainment level, sex and age (%)*. Available at: https://doi.org/10.2908/edat_lfs_9903

Eurostat (2024b). *Population on 1 January by age group and sex.* Available at: https://doi.org/10.2908/DEMO_PJANGROUP

Eurostat (2024c). *Demography of Europe - 2024 edition*. Available at: https://doi.org/10.2785/911441

Fast, E. and Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. *In Proceedings of the AAAI conference on artificial intelligence*, 31(1), pp. 963-969. https://doi.org/10.1609/aaai.v31i1.10635

Faul, F., Erdfelder, E., Lang, A.G. and Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), pp. 175-191. https://doi.org/10.3758/BF03193146

Field, A. and Hole, G. (2002). *How to design and report experiments.* SAGE Publications Ltd.

Field, A. P., Miles, J. and Field, Z. (2012). *Discovering statistics using R.* London: Sage.

Fox, J. (2015). *Applied regression analysis and generalized linear models.* London: Sage.

Francis, K. (2014). Timing: A Secret Weapon. *Qualtrics Blog*. 4 April. Available at: https://www.qualtrics.com/blog/timing-secret-weapon/ (Accessed: 2 June 2024).

Google (2023). Transforming the future of music creation. *Google DeepMind*. 16 November. Available at: https://bit.ly/3Rp2UPo (Accessed: 3 June 2024).

Greasley, A.E. and Lamont, A. (2011). Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae*, 15(1), pp. 45-71. https://doi.org/10.1177/10298649103934

Gregory, W. L. and Duran, A. (2001). Scenarios and acceptance of forecasts. *Principles of forecasting: A handbook for researchers and practitioners*, pp. 519-540. https://doi.org/10.1007/978-0-306-47630-3_23

Hadi, N. U., Abdullah, N. and Sentosa, I. (2016). An easy approach to exploratory factor analysis: Marketing perspective. *Journal of Educational and Social Research*, 6(1), pp. 215-223. https://doi.org/10.5901/jesr.2016.v6n1p215

Haque, M.U., Dharmadasa, I., Sworna, Z.T., Rajapakse, R.N. and Ahmad, H. (2022). 'I think this is the most disruptive technology': Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *arXiv preprint*. https://doi.org/10.48550/arXiv.2212.05856

Harkins, P. (2012). Extending the term: the Gowers Review and the campaign to increase the length of copyright in sound recordings. *Popular Music and Society*, 35(5), pp. 629-649. https://doi.org/10.1080/03007766.2012.709664

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6), pp. 836-850. https://doi.org/10.1016/j.cptl.2015.08.001

Hayes, A. F. and Rockwood, N. J. (2017). Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour Research and Therapy*, 98, pp. 39-57. https://doi.org/10.1016/j.brat.2016.11.001

Heerink, M., Kröse, B., Evers, V. and Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: The Almere model. *International Journal of Social Robotics*, 2(4), pp. 361- 375. https://doi.org/10.1007/s12369-010-0068-5

Henrich, J., Heine, S. J. and Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466 (29). https://doi.org/10.1038/466029a

Herndon, H. (2021). Holly +. *Mirror*. 13 July. Available at: https://bit.ly/4cBp6y7 (Accessed: 20 May 2024).

Hickey, M. (1999). Assessment Rubrics for Music Composition: Rubrics make evaluations concrete and objective, while providing students with detailed feedback and the skills to become sensitive music critics. *Music Educators Journal*, 85(4), 26–52. https://doi.org/10.2307/3399530

Hseih, V. (2024). Developing an Artificially Intelligent Voice: A Brief History of Text-to-Speech. *Deepgram*. 2 July. Available at: https://bit.ly/3xj3dUB (Accessed: 17 March 2024).

Holden, H. and Rada, R. (2011). Understanding the Influence of Perceived Usability and Technology Self-Efficacy on Teachers' Technology Acceptance, *Journal of Research on Technology in Education*, 43(4), pp. 343-367. https://doi.org/10.1080/15391523.2011.10782576

Hong, J. W. and Curran M. N. (2019). Artificial Intelligence, Artists, and Art: Attitudes Toward Artwork Produced by Humans vs. Artificial Intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(2), pp. 1-16. https://doi.org/10.1145/3326337

Hong, J.W. (2021). Artificial intelligence (AI), don't surprise me and stay in your lane: An experimental testing of perceiving humanlike performances of AI. *Human Behavior and Emerging Technologies*, 3, pp. 1023-1032. https://doi.org/10.1002/hbe2.292

Hong, J. W., Peng, Q. and Williams D. (2021). Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society* 23(7), pp. 1920-1935. https://doi.org/10.1177/1461444820925798

Hong, J. W., Fisher, K., Ha, J. and Zeng, Y. (2022). Human, I wrote a song for you: An experiment testing the influence of machines' attributes on the AI-composed music evaluation. *Computers in Human Behavior*, 131. https://doi.org/10.1016/j.chb.2022.107239

Hornaday, A. (2021). The controversy over Anthony Bourdain's deepfaked voice is a reminder that documentaries aren't journalism. *The Washington Post*. 19 July. Available at: https://wapo.st/45l7lQR (Accessed: 2 June 2024).

Huang, M. H. and Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21(2), pp. 155-172. https://doi.org/10.1177/1094670517752459

Ibrahim, S. (2023). AI-generated Drake song is an insult to the artistry of hip-hop. *The Washington Post*. 26 April. Available at: https://wapo.st/49vhoDV (Accessed: 10 February 2024).

Istók, E., Brattico, E., Jacobsen, T., Ritter, A. and Tervaniemi, M. (2013). 'I love Rock 'n'Roll' - Music genre preference modulates brain responses to music. *Biological Psychology*, 92(2), pp. 142-151. https://doi.org/10.1016/j.biopsycho.2012.11.005

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Janata, P. (1995). ERP measures assay the degree of expectancy violation of harmonic contexts in music. *Journal of Cognitive Neuroscience*, 7(2), pp. 153-164. https://doi.org/10.1162/jocn.1995.7.2.153

Jensen, M. L., Averbeck, J. M., Zhang, Z. and Wright, K. B. (2013). Credibility of anonymous online product reviews: A language expectancy perspective. *Journal of Management Information Systems*, 30(1), pp. 293-324. https://doi.org/10.2753/MIS0742-1222300109

Jhangiani, R. S., Chiang, I. C. A., Cuttler, C. and Leighton, D. C. (2019). *Research methods in psychology*. 4th edition. Kwantlen Polytechnic University. https://doi.org/10.17605/OSF.IO/HF7

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I. and Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, pp. 4485-4495. https://doi.org/10.48550/arXiv.1806.04558

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, pp. 31-36. https://doi.org/10.1007/BF02291575

Kalman, Y. M. and Sheizaf, R. (2010). Online Pauses and Silence: Chronemic Expectancy Violations in Written Computer-mediated Communication. *Communication Research*, 38(1), pp. 54-69. https://doi.org/10.1177/0093650210378229

Kang, H. (2021). Sample size determination and power analysis using the G* Power software. *Journal of educational evaluation for health professions*, 18(17). https://doi.org/10.3352/jeehp.2021.18.17

Kim, C. (2023). Can you copyright a voice? *Document Journal*. 18 April. Available at: https://bit.ly/AIdrake (Accessed: 10 February 2024).

Kock, F., Berbekova, A. and Assaf, A.G. (2021). Understanding and managing the threat of common method bias: Detection, prevention and control. *Tourism Management*, 86, p. 104330. https://doi.org/10.1016/j.tourman.2021.104330

Kwak, S. G. and Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*, 70(2), pp. 144-156. https://doi.org/10.4097/kjae.2017.70.2.144

Latikka, R., Bergdahl, J., Savela, N. and Oksanen, A. (2023). AI as an Artist? A Two-Wave Survey Study on Attitudes Toward Using Artificial Intelligence in Art. *Poetics*, 101. https://doi.org/10.1016/j.poetic.2023.101839

Lee, J., Choi, H. S., Jeon, C. B., Koo, J. and Lee, K. (2019). Adversarially trained end-to-end korean singing voice synthesis system. *arXiv preprint*. https://doi.org/10.48550/arXiv.1908.01919

Lee, Y., Kim, T. and Lee, S.Y. (2018). Voice imitating text-to-speech neural networks. *arXiv preprint*. https://doi.org/10.48550/arXiv.1806.00927

Lehmann, A. C. (1997). Research note: Affective responses to everyday life events and music listening. *Psychology of Music*, 25(1), pp. 84-90. https://doi.org/10.1177/0305735697251007

Liu, K. C., Wu, C. H., Tseng, S. Y. and Tsai, Y. T. (2015). Voice helper: A mobile assistive system for visually impaired persons. *In 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pp. 1400-1405. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.209

Lovett, M., Bajaba, S., Lovett, M. and Simmering, M.J. (2017). Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology*, 67(2), pp. 339-366. https://doi.org/10.1111/apps.12124

Machado, A. F. and Queiroz, M. (2010). 'Voice conversion: A critical survey'. *Proc. Sound and Music Computing Conference (SMC)*, pp. 1-8. https://doi.org/10.5281/ZENODO.849853

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90, pp. 46-60. https://doi.org/10.1016/j.futures.2017.03.006

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A. and Malik, H. (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, 53, pp. 3974-4026. https://doi.org/10.1007/s10489-022-03766-z

McKnight, P. E. and Najab, J. (2010). Mann-Whitney U Test. *The Corsini encyclopedia of psychology*. https://doi.org/10.1002/9780470479216.corpsy0524

Messingschlager, T. V. and Appel, M. (2023). Mind ascribed to AI and the appreciation of AI-generated art. *New media & society*. https://doi.org/10.1177/14614448231200248

Metz, C. (2023). The Secret Ingredient of ChatGPT Is Human Advice. *The New York Times*. 25 September. Available at: https://nyti.ms/3OIFnaL (Accessed: 10 February 2024).

Mulder, J., Ter Bogt, T. F., Raaijmakers, Q. A., Nic Gabhainn, S. and Sikkema, P. (2010). From death metal to R&B? Consistency of music preferences among Dutch adolescents and young adults. *Psychology of Music*, 38(1), pp. 67-83. https://doi.org/10.1177/0305735609104349

Myers, R. H. (1990). *Classical and modern regression with applications*. Belmont, CA: Duxbury Press.

Nachmani, E. and Wolf, L. (2019). Unsupervised singing voice conversion. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.06590

Nan, D., Lee, H., Kim, Y. and Kim, J. H. (2022). My video game console is so cool! A coolness theory-based model for intention to use video game consoles. *Technological Forecasting and Social Change*, 176. https://doi.org/10.1016/j.techfore.2021.121451

Novelli, N. and Proksch, S. (2022). Am I (deep) blue? Music-making AI and emotional awareness. *Frontiers in Neurorobotics*, 16, p. 897110. https://doi.org/10.3389/fnbot.2022.897110

Olshavsky, R. W. and Miller, J.A. (1972). Consumer expectations, product performance, and perceived product quality. *Journal of marketing research*, 9(1), pp. 19-21. https://doi.org/10.1177/002224377200900105

OpenAI (2024). Navigating the Challenges and Opportunities of Synthetic Voices. *OpenAI*. 29 March. Available at: https://bit.ly/OpenArtInT (Accessed: 8 June 2024).

Oppenheimer, D. M., Meyvis, T. and Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4), pp. 867-872. https://doi.org/10.1016/j.jesp.2009.03.009

Oramas, S., Nieto, O., Barbieri, F. and Serra, X. (2017). Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint*. https://doi.org/10.48550/arXiv.1707.04916

Osborne, J.W. and Overbay, A. (2019). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1), p. 6. https://doi.org/10.7275/qf69-7k43

Pachet, F. and Cazaly, D. (2000). A taxonomy of musical genres. *In RIAO 00: Content-Based Multimedia Information Access*, 2, pp. 1238-1245. https://dl.acm.org/doi/10.5555/2856151.2856177

Palan, S. and Schitter, C. (2018). Prolific.ac - A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, pp. 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S. and Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, pp. 153-163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z. and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), pp. 1643-1662. https://doi.org/10.3758/s13428-021-01694-3

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J. and Miller, J. (2017). Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1710.07654

Prolific (2023). What are the advantages and limitations of an online sample? *Prolific*. 25 October. Available at: https://bit.ly/prolifc (Accessed: 10 June 2024).

Prolific (2024a). What is Prolific? *Prolific*. 1 May. Available at: https://bit.ly/3xtq807 (Accessed: 2 June 2024).

Prolific (2024b). Prolific's Attention and Comprehension Check Policy. *Prolific*. 8 May. Available at: https://bit.ly/3zrLZFP (Accessed: 10 June 2024).

Prolific (n.d.). Prolific vs. MTurk. *Prolific*. Available at: https://www.prolific.com/prolific-vs-mturk (Accessed: 10 June 2024).

Qualtrics (n.d.a). Randomizer. *Qualtrics Support*. Available at: https://bit.ly/4cq3ftx (Accessed: 4 June 2024).

Qualtrics (n.d.b). Meta Info Question. *Qualtrics Support*. https://bit.ly/4bqG5Sx (Accessed: 2 June 2024).

Regev, M. (1994). Producing artistic value: The case of rock music. *The Sociological Quarterly*, 35(1), pp. 85-102. https://doi.org/10.1111/j.1533-8525.1994.tb00400.x

Regorz, A. (2021). R - Moderation Analysis with PROCESS Model 1. *Regorz Statistics*. 31 March. Available at: http://www.regorz-statistik.de/en/moderation_process_for_r.html (Accessed: 5 July 2024)

Reid, A. (2001). Variation in the Ways that Instrumental and Vocal Students Experience Learning Music. *Music Education Research*, 3(1), pp. 25-40. https://doi.org/10.1080/14613800020029932

Rentfrow, P. J., Goldberg, L.R. and Levitin, D.J. (2011). The structure of musical preferences: a five-factor model. *Journal of Personality and Social Psychology*, 100(6), pp. 1139-1157. https://doi.org/10.1037/a0022406

Rentfrow, P. J. and Gosling, S.D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6), pp. 1236-1256. http://dx.doi.org/10.1016/j.biopsycho.2012.11.005

Roehm, M. L. (2001). Instrumental vs. vocal versions of popular music in advertising. *Journal of Advertising Research*, 41(3), pp. 49-58. http://dx.doi.org/10.2501/JAR-41-3-49-58

Samuels, P. (2016). Advice on exploratory factor analysis. *Technical Report*. Research Gate. https://doi.org/10.13140/RG.2.1.5013.9766

Schäfer, T. and Sedlmeier, P. (2010). What makes us like music? Determinants of music preference. *Psychology of Aesthetics, Creativity, and the Arts*, 4(4), pp. 223-234. https://doi.org/10.1037/a0018374

Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K. and Belfi, A. M. (2023). AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied*, 29(3), pp. 676-692. https://doi.org/10.1037/xap0000447

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R. and Saurous, R. A. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779-4783. https://doi.org/10.1109/ICASSP.2018.8461368

Shevlin, A. and Shan L. (2019). Tencent in Talks for Stake in Record Label of Ariana Grande, Queen. *The Wall Street Journal*. 6 August. Available at: https://on.wsj.com/3whA6Rh (Accessed: 10 February 2024).

Shrivastava, R. (2023). Keep Your Paws Off My Voice': Voice Actors Worry Generative AI Will Steal Their Livelihoods. *Forbes*. 9 October. Available at: https://bit.ly/AIVoiceActors (Accessed: 3 June 2024).

Sisman, B., Yamagishi, J., King, S. and Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp. 132-157. https://doi.org/10.1109/TASLP.2020.3038524

Snapes, L. (2023). AI song featuring fake Drake and Weeknd vocals pulled from streaming services. *The Guardian*. 18 April. Available at: https://bit.ly/theguardianfakedrake (Accessed: 10 February 2024).

Spence, P. R., Edwards, C., Edwards, A. and Lin, X. (2019). Testing the machine heuristic: Robots and suspicion in news broadcasts. *In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 568-569. https://doi.org/10.1109/HRI.2019.8673108

Spotify (2024). Our Annual Music Economics Report. *Loud and Clear*. Available at: https://loudandclear.byspotify.com (Accessed: 25 June 2024).

Squires, L. (2019). Genre and linguistic expectation shift: Evidence from pop song lyrics. *Language in Society*, 48(1), pp. 1-30. https://doi.org/10.1017/S0047404518001112

Steinbeis, N., Koelsch, S. and Sloboda J.A. (2006). The role of harmonic expectancy violations in musical emotions: evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience*, 18(8), pp. 1380-1393. https://doi.org/10.1162/jocn.2006.18.8.1380

Steinbrecher, B. (2021). Mainstream popular music research: a musical update. *Popular Music*, 40(3-4), pp. 406-427. https://doi.org/10.1017/S0261143021000568

Stewart, J. Q. (1922). An electrical analogue of the vocal organs. *Nature*, 110(2757), pp. 311-312. https://doi.org/10.1038/110311a0

Story, B. H. (2019). History of speech synthesis. In Katz, W. and Assmann, P. (eds.) *The Routledge Handbook of Phonetics*. Routledge, pp. 9-33.

Stupp, C. (2019). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *The Wall Street Journal.* 30 August. Available at: https://on.wsj.com/3Vm7pvl (Accessed: 2 June 2024).

Sturm, B. L. T., Iglesias, M., Ben-Tal, O., Miron, M. and Gómez, E. (2019). Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. *Arts*, 8(3). https://doi.org/10.3390/arts8030115

Sun, D., Wang, H. and Xiong, J. (2023). Would You Like to Listen to My Music, My Friend? An Experiment on AI Musicians. *International Journal of Human-Computer Interaction*, pp. 1-11. https://doi.org/10.1080/10447318.2023.2181872

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in science education*, 48, pp. 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press.

Tigre Moura, F. and Maw, C. (2021). Artificial intelligence became Beethoven: How do listeners and music professionals perceive artificially composed music? *Journal of Consumer Marketing*, 38(2), pp. 137-146. https://doi.org/10.1108/JCM-02-2020-3671

Trizano-Hermosilla, I. and Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Frontiers in Psychology*, 7(769). https://doi.org/10.3389/fpsyg.2016.00769

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5), pp. 293-302. https://doi.org/10.1109/TSA.2002.800560

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint* https://doi.org/10.48550/arXiv.1609.03499

Veaux, C., Yamagishi, J. and King, S. (2013). Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction. *In Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pp. 107-111. https://doi.org/10.1250/ast.33.1

Vilenkin, R. (2023). How AI Voice Cloning Technology Helps Film Studios Restore Old Movies. *Respeecher*. 28 March. Available at: https://bit.ly/cloningactors (Accessed: 3 June 2024).

Waddell, T. (2018). A robot wrote this? How perceived machine authorship affects news credibility. *Digital Journalism*, 6(2), pp. 236-255. https://doi.org/10.1080/21670811.2017.1384319

Walczyna, T. and Piotrowski, Z. (2023). Overview of voice conversion methods based on deep learning. *Applied Sciences*, 13(5), p. 3100. https://doi.org/10.3390/app13053100

Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L. and Liu, Y. (2020). Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. *In Proceedings of the 28th ACM international conference on multimedia*, pp. 1207-1216. https://doi.org/10.1145/3394171.3413716

Weitzman, C. (2022). A short history of text to speech. *Speechify*. 27 June. Available at: https://bit.ly/3VGNJE3 (Accessed: 17 March 2024).

Weitzman, C. (2023a). Voice Cloning for Music. *Speechify*. 17 August. Available at: https://bit.ly/voicecloningmusic (Accessed: 10 February 2024).

Weitzman, C. (2023b). Voice Cloning for Singing. *Speechify*. 19 August. Available at: https://bit.ly/3XfDQy9 (Accessed: 25 May 2024).

Wiggers, K. (2024). People are using AI music generators to create hateful songs. *TechCrunch*. Available at: https://tcrn.ch/4ehGPvZ (Accessed: 7 June 2024).

Wirtz, B.W., Weyerer, J.C. and Geyer, C. (2018). Artificial intelligence and the public sector - applications and challenges. *International Journal of Public Administration*, 42(7), pp. 596-615. https://doi.org/10.1080/01900692.2018.1498103

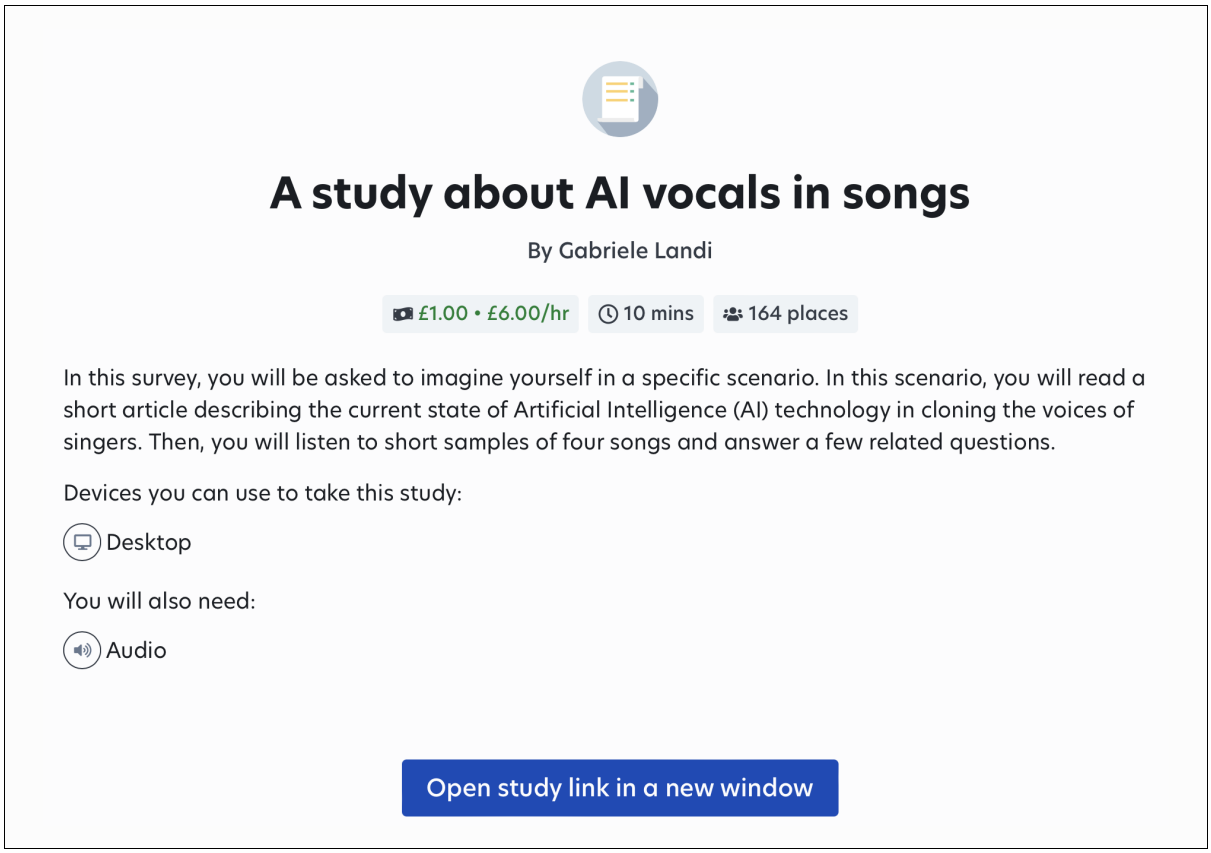
Zdaniuk, B. (2014). *Ordinary Least-Squares (OLS) model.* In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_2008

Zhang, C. and Conrad, F. G. (2014). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8(2), pp. 127-135. https://doi.org/10.18148/srm/2014.v8i2.5453

Zhang, R., Mao, X., Li, L., Jiang, L., Chen, L., Hu, Z., Xi, Y., Fan, C. and Huang, M. (2022). Youling: an AI-assisted lyrics creation system. *arXiv preprint*. https://doi.org/10.48550/arXiv.2201.06724

# 9 Appendix

## 9.1 Appendix A




*Pre-survey briefing (a)*

**Welcome!**

In this survey, you will be asked to imagine yourself in a specific scenario. In this scenario, you will read a short article describing the current state of Artificial Intelligence (AI) technology in cloning the voices of singers. Then, you will listen to short samples of four songs and answer a few related questions.

Please take your time to read the article carefully, listen to the song samples with a pair of headphones, and provide thoughtful responses to the survey questions.

**Important Information:**

- **Estimated Time**: This survey should take you around 10 minutes to complete.
- **Anonymity**: The survey is anonymous. Your name and identifying information will not be connected to your answers in any way. The only information visible to the researcher will be your Prolific ID, which is required for compensation purposes.
- **Data**: All data will be stored securely and will only be accessible to the research team. The data collected will be used exclusively for research purposes.
- **Voluntary Participation**: Participation in this study is completely voluntary. You are free to decline to participate or to end your participation at any time for any reason. If you wish to withdraw from the study at any point, simply close your browser window. Your responses will not be recorded, and you will not be compensated.
- **Recommendations**: For the best listening experience, headphones are required when listening to the song sample. It is recommended to have a stable internet connection.
- **Attention Checks**: Please be aware that there are randomly distributed attention checks throughout the survey. Pay close attention to all questions; if you fail to appropriately respond to the attention checks, you will not be compensated.
- **One-Time Participation**: You can only complete this survey once. Duplicate responses will not be compensated.
- **No Backtracking**: You will not be allowed to go back to previously answered questions.

**Contact Information:**

This study is conducted by Gabriele Landi, a Master's student at Rotterdam School of Management. If you have any questions about this study, you may contact the researcher at gabriel.landi@outlook.it.

Thank you again for your valuable participation. Your input is greatly appreciated!

By proceeding with this survey, you confirm that you have read and understood the information provided and agree to participate in this study.

*Pre-survey briefing (b)*

What is your Prolific ID? *Please note that this response should auto-fill with the correct ID*

*Prolific ID*

## 9.2 Appendix B

If you are unfamiliar with any of the terminologies below, please take a moment to read the definitions provided. Once you are ready, please click the arrow below to continue.

- **VOCALS**: the singing in a piece of music. It refers to a section of music that's sung, rather than played on an instrument.
- **AI-CLONED VOCALS**: the singing in a piece of music that has been generated or replicated using artificial intelligence technology. This involves creating a synthetic voice that mimics a specific real human singer's voice.

*Definitions*

## 9.3 Appendix C

Please indicate the extent to which you agree or disagree with the following statements about Artificial Intelligence (AI) in the creative industry (music, visual arts, etc...).

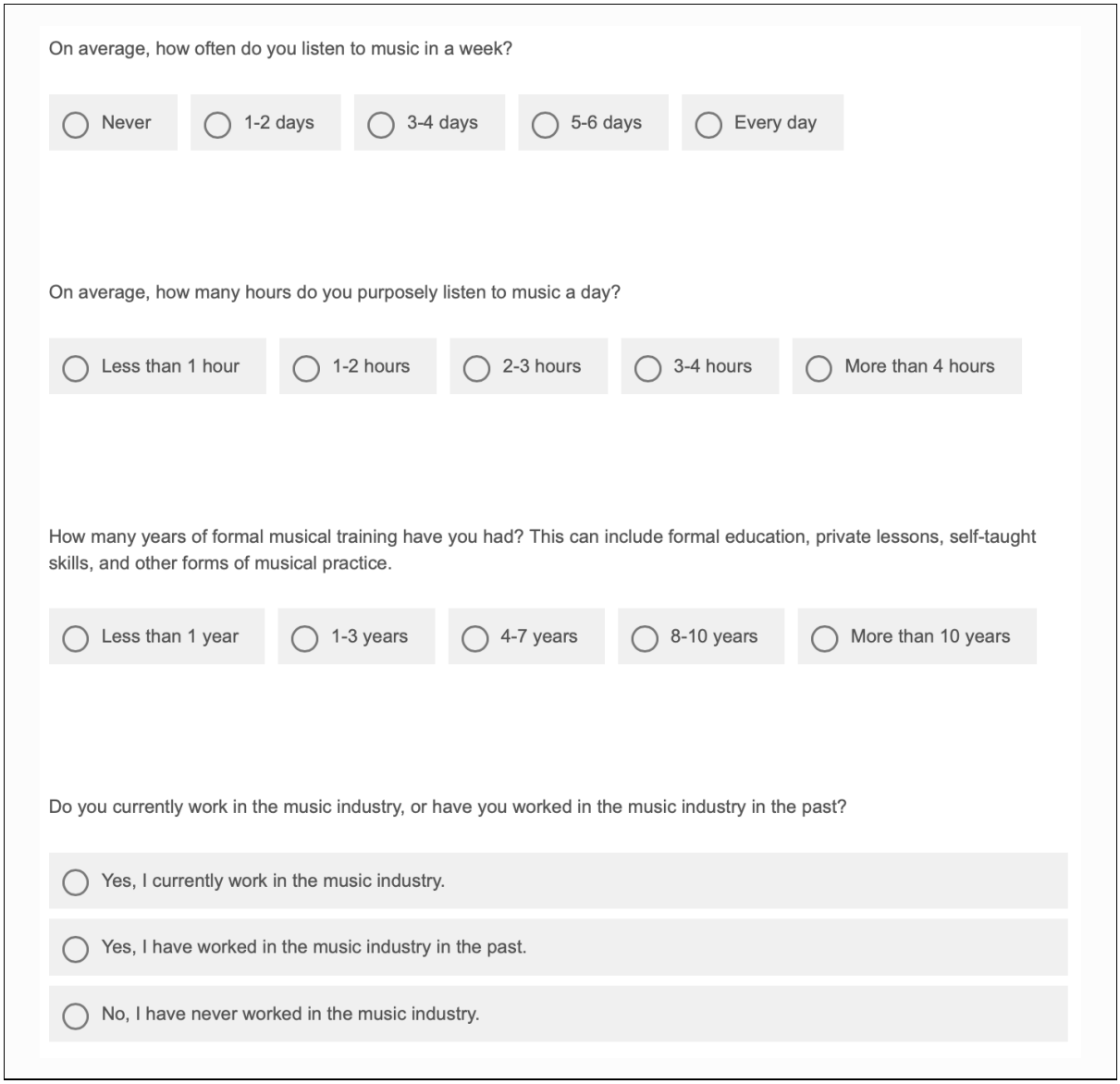
| | Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|---|
| I believe AI can make something new by itself. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I think AI can be creative on its own. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Products developed by AI should be respected as creative works. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

*Attitudes towards creative AI*

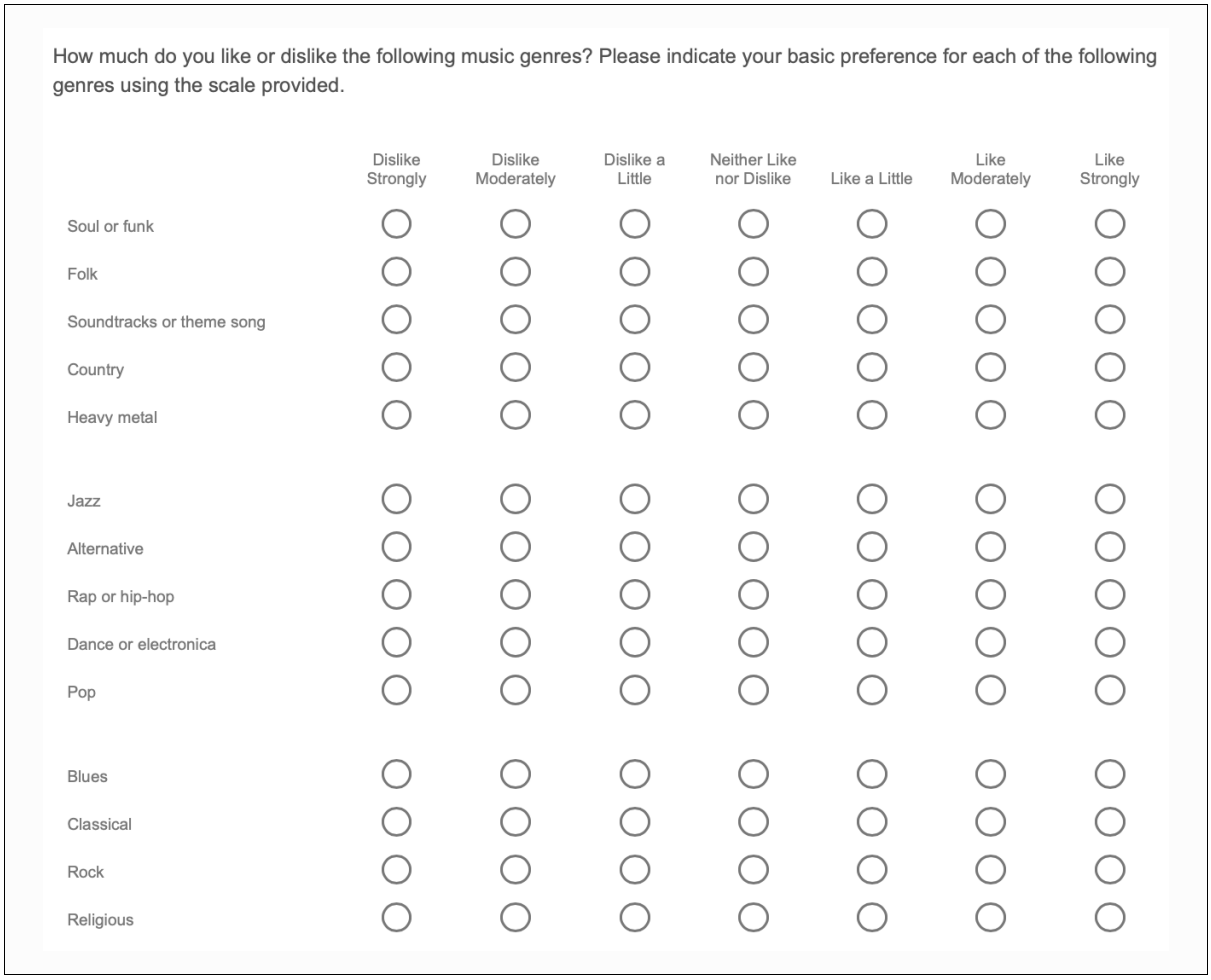
Please indicate the extent to which you agree or disagree with the following statements about Artificial Intelligence (AI) .

| | Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|---|
| I think people should accept the AI that clones vocals as a singer. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the AI that clones vocals should be regarded as a singer. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I think the AI that clones vocals qualifies to become a singer. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

*Acceptance of AI as a singer*




*IML and musical training*



How much do you like or dislike the following music genres? Please indicate your basic preference for each of the following genres using the scale provided.

| | Dislike Strongly | Dislike Moderately | Dislike a Little | Neither Like nor Dislike | Like a Little | Like Moderately | Like Strongly |
|---|---|---|---|---|---|---|---|
| Soul or funk | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Folk | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Soundtracks or theme song | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Country | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Heavy metal | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Jazz | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Alternative | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Rap or hip-hop | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Dance or electronica | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Pop | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Blues | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Classical | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Rock | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Religious | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

*Music genre preferences*

## 9.4 Appendix D

**Imagine that you are hearing about AI-cloned vocals in songs for the first time.** You are curious to know more about the technology, so you do some research and come across this article featured in a recent issue of SoundWave Monthly, a respected magazine in AI technology. Please take your time to read it carefully.

SOUNDWAVE MONTHLY

**The Remarkable Quality of Songs with AI-Cloned Vocals**

by ***Isabel Woods***
*Isabel Woods writes about artificial intelligence and other cutting-edge technologies.*

*May 15, 2024*

[..] AI technology used to clone vocals in songs has made great progress, resulting in songs of excellent quality that sound just as good as those sung by real people. This cutting-edge technology uses sophisticated algorithms and machine learning techniques to replicate accurately the nuances and emotional depth of human voices.

Recent songs featuring AI-cloned vocals have collected very positive reviews. '*The sound quality is extremely realistic*' remarks John Peterson, a seasoned music critic. '*The current AI technology used for cloning vocals is very advanced*!' says Dr. Emily Hartman, a researcher in music technology.

Industry peers are also enthusiastic about the quality of AI-cloned vocals. '*The AI-cloned vocals have the warmth and authenticity of real human singers*' says Grammy-winning producer Sarah Lee. Listeners have shared similar sentiments, '*I was impressed with the sound*,' comments Michael, '*I didn't notice any differences from a human song, WOW!*'

[..] The excitement around AI-cloned vocals is well-founded, leaving many optimistic about its ability to deliver high-quality vocals.

*Article 1 - Positive*

What does the article discuss?

- ◯ The history of classical music
- ◯ The use of AI in movies
- ◯ The use of AI in instrumental music
- ◯ The high-quality of AI-cloned vocals in songs

*Article 1 - Check*

**Imagine that you are hearing about AI-cloned vocals in songs for the first time.** You are curious to know more about the technology, so you do some research and come across this article featured in a recent issue of SoundWave Monthly, a respected magazine in AI technology. Please take your time to read it carefully.

SOUNDWAVE
MONTHLY

**The Poor Quality of Songs with AI-Cloned Vocals**

by ***Isabel Woods***
*Isabel Woods writes about artificial intelligence and other cutting-edge technologies.*

*May 15, 2024*

[..] Despite the hype, AI technology used to clone vocals in songs faces several limitations, leading to songs that sound extremely robotic, artificial, and lacking quality. This technology uses basic algorithms that still fail to capture the characteristics, subtle nuances and emotional depth of human voices.

Recent songs featuring AI-cloned vocals have collected negative reviews. '***The sound quality is awful, I cannot even understand the lyrics***' remarks John Peterson, a seasoned music critic. '***The technology is extremely rudimentary***,' says Dr Emily Hartman, a researcher in music technology.

Industry peers have also voiced their concerns. '***The AI-cloned vocals sound like bad robots, lacking any authenticity of real human singers***,' says Grammy-winning producer Sarah Lee. Listeners have shared similar sentiments, '***I was disappointed with the sound***,' comments Michael, '***It just doesn't sound human, I cannot even understand what it has been sung***.'

[..] Despite the initial excitement, AI-cloned vocals seem to be falling flat, leaving many to wonder if this technology will ever be able to reach the desired quality.

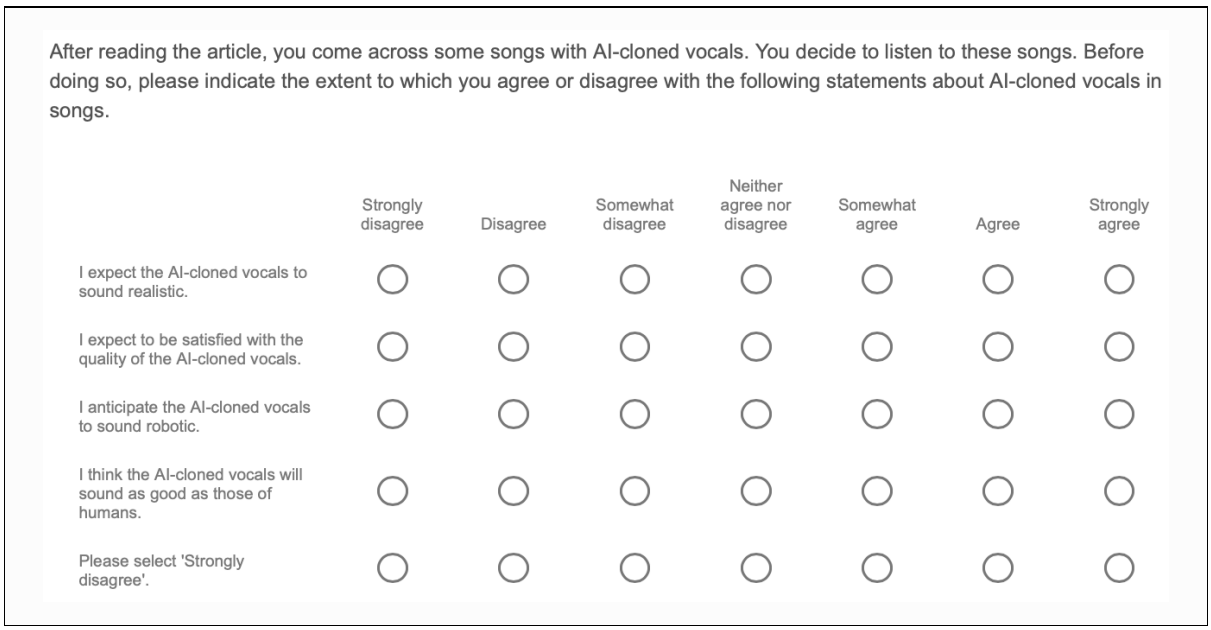
*Article 2 - Negative*

What does the article discuss?

- ◯ The use of AI in movies
- ◯ The use of AI in instrumental music
- ◯ The history of classical music
- ◯ The low-quality of AI-cloned vocals in songs

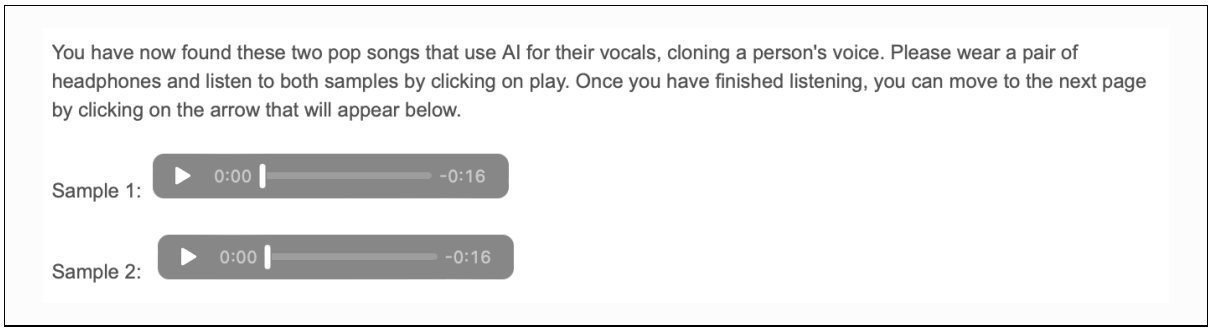
*Article 2 - Check*

## 9.5 Appendix E

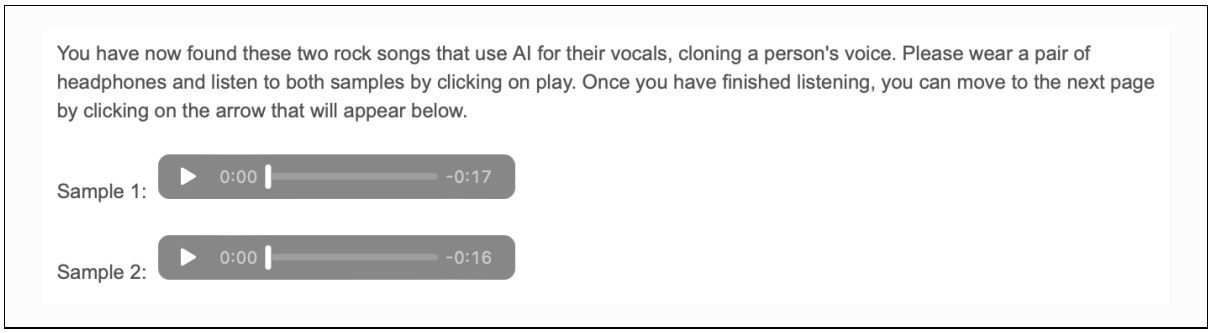



*Expectations manipulation check*

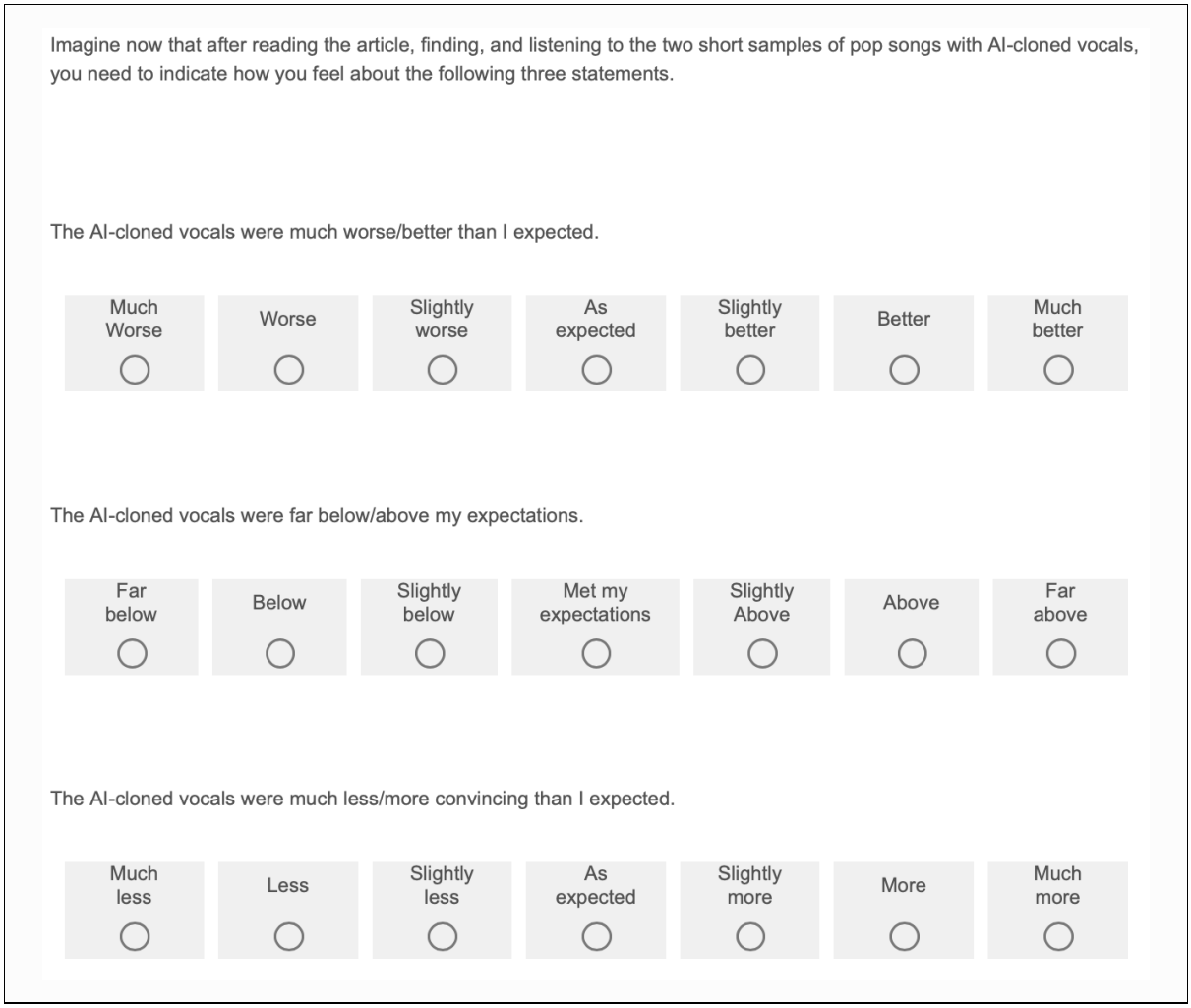
## 9.6 Appendix F




*Exposure to samples - Pop*




*Exposure to Samples - Rock*



Imagine now that after reading the article, finding, and listening to the two short samples of pop songs with AI-cloned vocals, you need to indicate how you feel about the following three statements.

The AI-cloned vocals were much worse/better than I expected.

| Much Worse | Worse | Slightly worse | As expected | Slightly better | Better | Much better |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

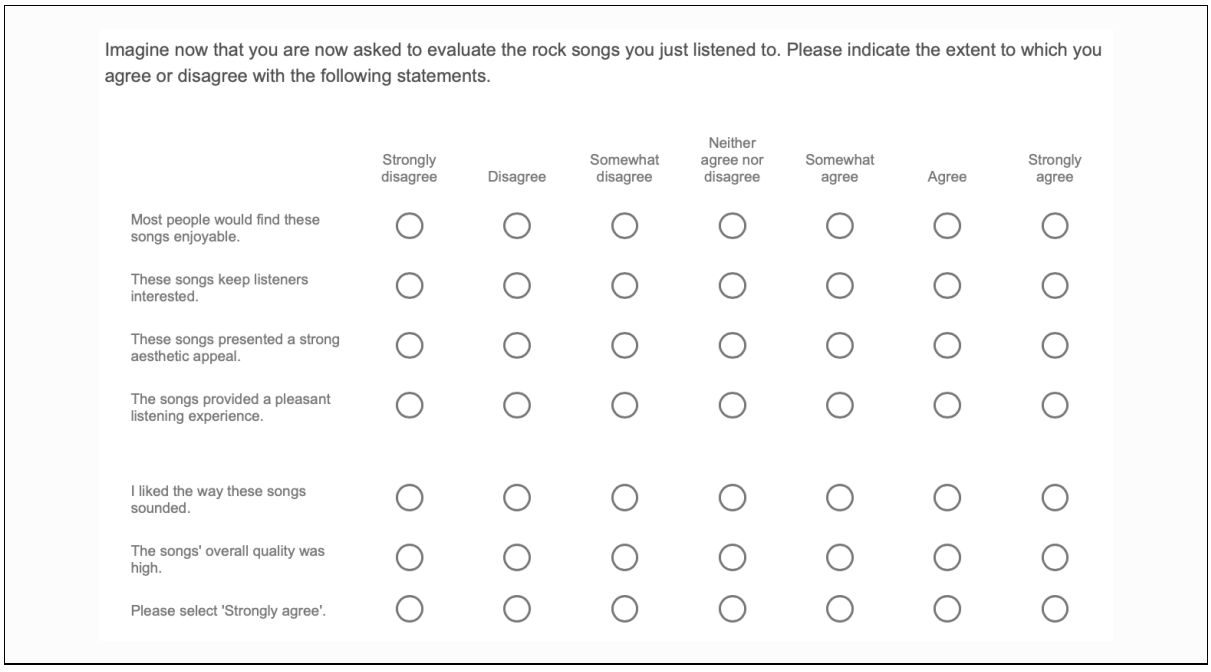
The AI-cloned vocals were far below/above my expectations.

| Far below | Below | Slightly below | Met my expectations | Slightly Above | Above | Far above |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

The AI-cloned vocals were much less/more convincing than I expected.

| Much less | Less | Slightly less | As expected | Slightly more | More | Much more |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

*Expectancy violation*

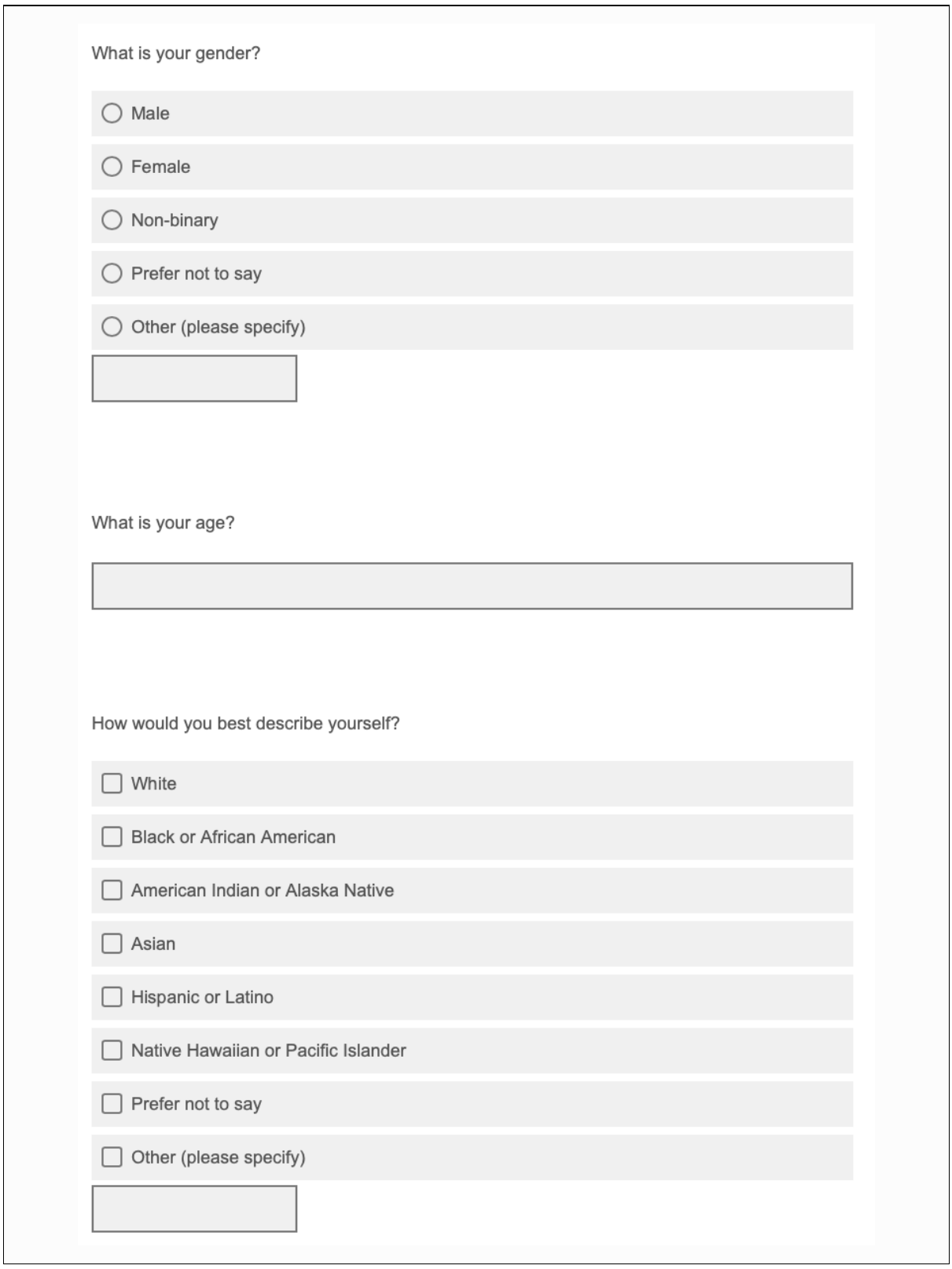


Imagine now that you are now asked to evaluate the rock songs you just listened to. Please indicate the extent to which you agree or disagree with the following statements.

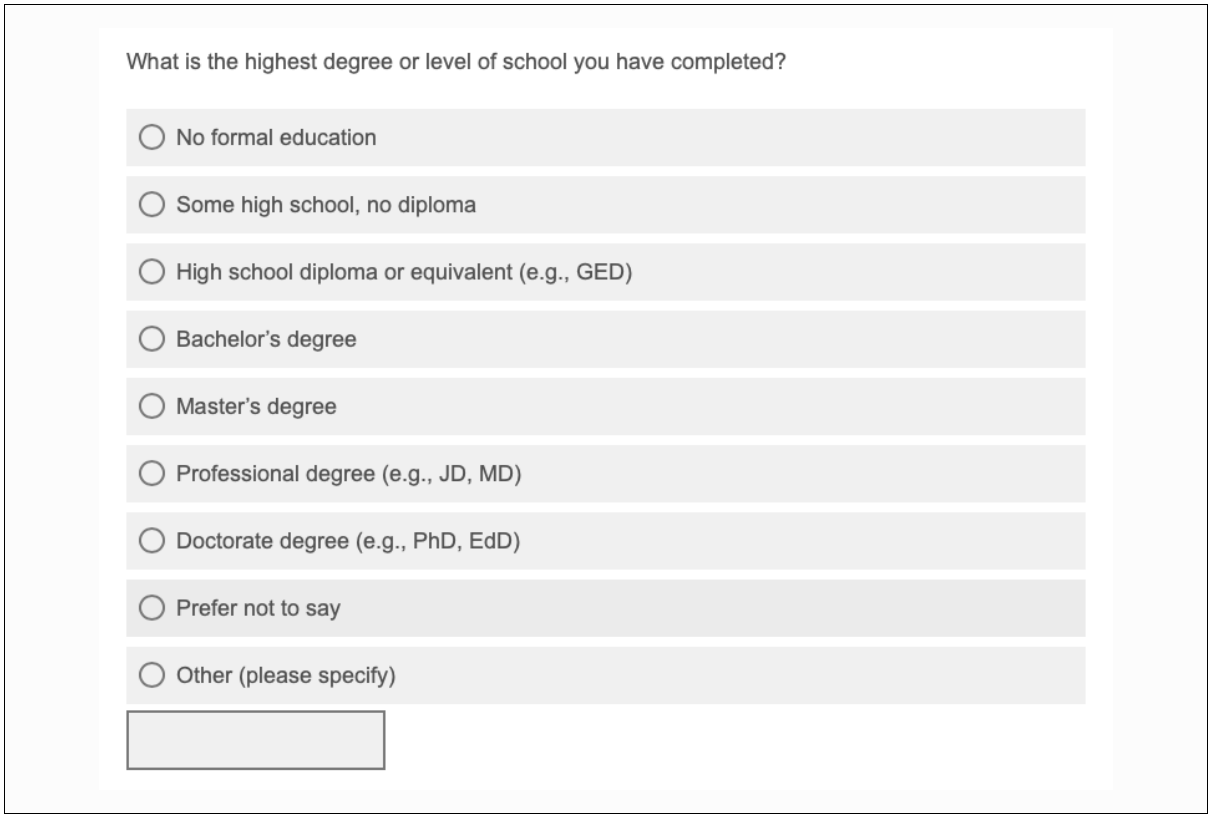
| | Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|---|
| Most people would find these songs enjoyable. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| These songs keep listeners interested. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| These songs presented a strong aesthetic appeal. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The songs provided a pleasant listening experience. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| I liked the way these songs sounded. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The songs' overall quality was high. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Please select 'Strongly agree'. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

*Samples evaluation*

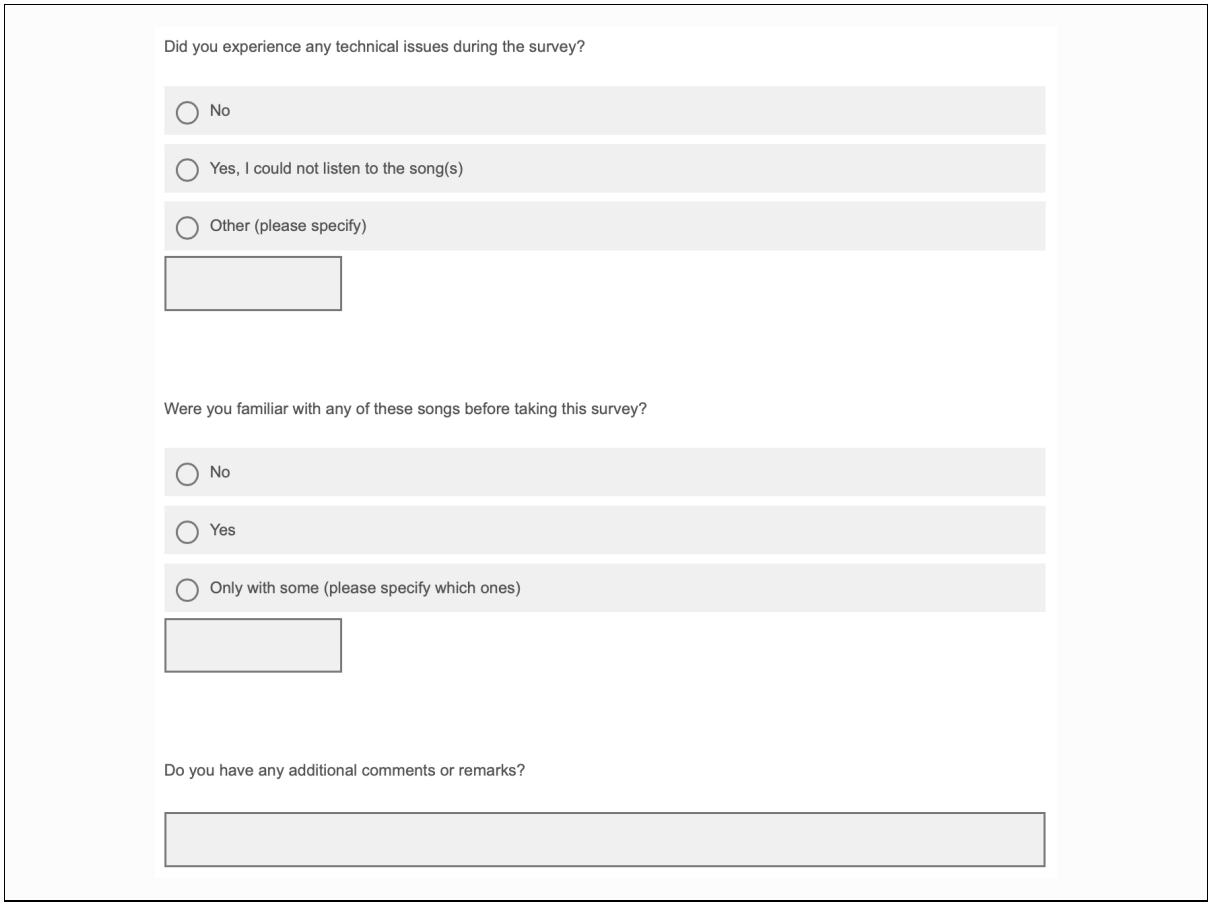
## 9.7 Appendix G



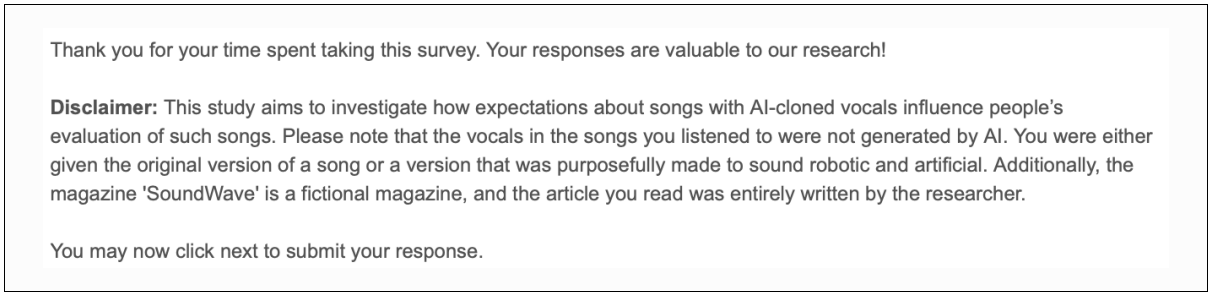

*Demographics (a)*




*Demographics (b)*




*Technical issues & remarks*


Thank you for your time spent taking this survey. Your responses are valuable to our research!

**Disclaimer:** This study aims to investigate how expectations about songs with AI-cloned vocals influence people's evaluation of such songs. Please note that the vocals in the songs you listened to were not generated by AI. You were either given the original version of a song or a version that was purposefully made to sound robotic and artificial. Additionally, the magazine 'SoundWave' is a fictional magazine, and the article you read was entirely written by the researcher.

You may now click next to submit your response.

*Disclaimer*

## 9.8 Appendix H

We selected two songs for the pop genre and two songs for the rock genre and created 15-second samples. The songs are performed in English by artists typically performing in these genres. Our choices were further validated by experts in the field of music. The songs chosen were among the least popular of these artists to further reduce the likelihood of participants being familiar with them.

1. Female Pop - 'I Feel it in the Earth' by Elisa
2. Male Pop - 'Close my Eyes' by Lorenzo Fragola
3. Female Rock - 'Something Sweet' by Madison Beer
4. Male Rock - 'NVR 4EVR' by Death From Above 1979

We extracted the vocals from the songs using an online tool available at https://www.lalal.ai/. Using GarageBand, a software application by Apple for macOS, we purposely created lower-quality versions of the vocals by selecting the Robot Vocal option. The figure below illustrates this process. In this way, we were able to confidently create four additional samples of lower vocal quality.


*Vocal manipulation - Example*

## 9.9 Appendix I

*Measurements Constructs (a)*

| | |
|---|---|
| **1. Attitudes Towards Creative AI** (Hong et al., 2021) | |
| **ATC1** | I think AI can be creative on its own. |
| **ATC2** | I believe AI can make something new by itself. |
| **ATC3** | Products developed by AI should be respected as creative works. |
| **2. Acceptance of AI as a Singer** (adapted from Hong et al., 2022) | |
| **AS1** | I think the AI that clones vocals should be regarded as a singer. |
| **AS2** | I think the AI that clones vocals qualifies to become a singer. |
| **AS3** | I think people should accept the AI that clones vocals as a singer. |
| **3. Short Test of Music Preferences (STOMP)** (Rentfrow and Gosling, 2003) | |
| Alternative, Blues, Classical, Country, Dance or electronica, Folk, Heavy metal, Jazz, Rap or Hip-Hop, Religious, Soul or funk, Soundtracks or theme song | |
| **4. Index of Music Listening** (Chin and Rickard, 2012) | |
| **IML1** | On average, how often do you listen to music in a week? |
| **IML2** | On average, how many hours do you purposely listen to music a day? |
| **5. Formal Musical Training** (adapted from Chin and Rickard, 2012) | |
| **FMT1** | How many years of formal music training have you had? This can include formal education, private lessons, self-taught skills, and other forms of musical practice. |

Notes: All items in 1 and 2 were measured on a 7-point Likert-type scale, ranging from strongly disagree to strongly agree. All items in 3 were measured on a 7-point Likert-type scale, ranging from dislike strongly to like strongly. All items in 4 and 5 were measured on a multiple-choice scale with various frequency options.

**6. Expectations of AI-cloned vocals** (adapted from Olshavsky and Miller, 1972)

| | |
|---|---|
| **E1** | I expect to be satisfied with the quality of the AI-cloned vocals. |
| **E2** | I expect the AI-cloned vocals to sound realistic. |
| **E3** | I anticipate the AI-cloned vocals to sound robotic. |
| **E4** | I think the AI-cloned vocals will sound as good as those of humans. |

**7. Expectancy Violation** (adapted from Messingschlager and Appel, 2023)

| | |
|---|---|
| **EV1** | The AI-cloned vocals were much worse/better than I expected. |
| **EV2** | The AI-cloned vocals were far below/above my expectations. |
| **EV3** | The AI-cloned vocals were much less/more convincing than I expected. |

**8. Songs Evaluation** (adapted from Hickey, 1999; Hong et al., 2021)

| | |
|---|---|
| **SE1** | Many listeners would enjoy this song. |
| **SE2** | This song keeps listeners interested. |
| **SE3** | The song presented a strong aesthetic appeal. |
| **SE4** | The song provided a pleasant listening experience. |
| **SE5** | I liked the way the song sounded. |
| **SE6** | The song's overall quality was high. |

Notes: All items in 6 and 8 were measured on a 7-point Likert-type scale, ranging from strongly disagree to strongly agree. All items in 7 were measured on a 7-point bipolar scale.

## 9.10 Appendix J

*Demographic Statistics*

| | Count | Percentage |
|---|---|---|
| **Sample Size** | 158 | |
| **Gender** | | |
| Male | 91 | 0.58 |
| Female | 64 | 0.41 |
| Non-binary | 3 | 0.02 |
| **Age** | | |
| 18-24 years | 37 | 0.23 |
| 25-34 years | 72 | 0.46 |
| 35-44 years | 36 | 0.23 |
| 45-54 years | 10 | 0.06 |
| 55+ years | 3 | 0.02 |
| **Education** | | |
| High school | 42 | 0.27 |
| Bachelor's degree | 57 | 0.36 |
| Master's degree | 46 | 0.29 |
| Professional degree | 5 | 0.03 |
| Doctorate degree | 4 | 0.03 |
| Other | 4 | 0.03 |
| **Ethnicity** | | |
| White | 151 | 0.96 |
| Asian | 5 | 0.03 |
| Hispanic or Latino | 2 | 0.01 |
| **Musical Training** | | |
| Less than 1 year | 85 | 0.54 |
| 1-3 years | 38 | 0.24 |
| 4-7 years | 18 | 0.11 |
| 8-10 years | 12 | 0.08 |
| More than 10 years | 5 | 0.03 |
| **Index of Music Listening** | | |
| 2-5 | 45 | 0.28 |
| 6-10 | 53 | 0.36 |
| 12-16 | 23 | 0.15 |
| 20-25 | 37 | 0.23 |
| **Nationality** | | |
| European Union | 151 | 0.96 |
| Other | 7 | 0.04 |
| **Country of residence** | | |
| European Union | 158 | 1.00 |

Note: Values are rounded to the nearest cent.

## 9.11 Appendix K

*KMO Measure of Sampling Adequacy*

| | MSA (Pop) | MSA (Rock) |
|---|---|---|
| **Overall** | 0.899 | 0.898 |
| **ATC1** | 0.801 | 0.800 |
| **ATC2** | 0.854 | 0.869 |
| **ATC3** | 0.787 | 0.789 |
| **AS1** | 0.789 | 0.796 |
| **AS2** | 0.791 | 0.785 |
| **AS3** | 0.876 | 0.878 |
| **E1** | 0.829 | 0.822 |
| **E2** | 0.869 | 0.856 |
| **E3** | 0.823 | 0.796 |
| **E4** | 0.851 | 0.840 |
| **EV1** | 0.916 | 0.904 |
| **EV2** | 0.901 | 0.893 |
| **EV3** | 0.919 | 0.949 |
| **SE1** | 0.941 | 0.951 |
| **SE2** | 0.944 | 0.953 |
| **SE3** | 0.954 | 0.943 |
| **SE4** | 0.939 | 0.928 |
| **SE5** | 0.933 | 0.935 |
| **SE6** | 0.962 | 0.952 |

Note: Values are rounded to the nearest thousandth.

*Bartlett's Test of Sphericity*

| | $\chi^2$ | df | p |
|---|---|---|---|
| **Pop** | 4587 | 171 | <.001 |
| **Rock** | 4400 | 171 | <.001 |

Note: Values are rounded to the nearest thousandth.

*Inter-Factor Correlations (Pop)*

| | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **1** | - | 0.360 | 0.182 | 0.050 | 0.142 |
| **2** | - | - | 0.208 | 0.025 | 0.112 |
| **3** | - | - | - | -0.166 | 0.129 |
| **4** | - | - | - | - | 0.772 |
| **5** | - | - | - | - | - |

Note: Values are rounded to the nearest thousandth.

*Exploratory Factor Analysis (rock)*

| | **Factor** | | | | | |
|---|---|---|---|---|---|---|
| **Items** | **1** | **2** | **3** | **4** | **5** | **Uniqueness** |
| ATC1 | 0.846 | | | | | 0.245 |
| ATC2 | 0.727 | | | | | 0.403 |
| ATC3 | 0.887 | | | | | 0.216 |
| AS1 | | 0.899 | | | | 0.174 |
| AS2 | | 0.924 | | | | 0.161 |
| AS3 | | 0.778 | | | | 0.326 |
| E1 | | | 0.854 | | | 0.257 |
| E2 | | | 0.768 | | | 0.261 |
| E3 | | | 0.890 | | | 0.273 |
| E4 | | | 0.848 | | | 0.230 |
| EV1 | | | | 0.937 | | 0.059 |
| EV2 | | | | 0.957 | | 0.051 |
| EV3 | | | | 0.858 | | 0.129 |
| SE1 | | | | | 0.673 | 0.242 |
| SE2 | | | | | 0.835 | 0.168 |
| SE3 | | | | | 0.935 | 0.212 |
| SE4 | | | | | 0.875 | 0.084 |
| SE5 | | | | | 0.948 | 0.138 |
| SE6 | | | | | 0.869 | 0.143 |

Note: The ‘Principal axis factoring’ extraction method was combined with an ’oblimin’ rotation. Values are rounded to the nearest thousandth.

*Inter-Factor Correlations (rock)*

| | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **1** | - | 0.325 | 0.182 | 0.018 | 0.129 |
| **2** | - | - | 0.196 | -0.013 | 0.106 |
| **3** | - | - | - | -0.207 | 0.057 |
| **4** | - | - | - | - | 0.800 |
| **5** | - | - | - | - | - |

Note: Values are rounded to the nearest thousandth.

## 9.12 Appendix L

*Normality Test (Shapiro-Wilk) – Article 1 Group vs. Article 2 Group*

| **Experimental Condition** | **W** | **p** |
|---|---|---|
| Age | 0.912 | <.001 |
| IML | 0.916 | <.001 |
| Musical Training | 0.731 | <.001 |
| Music Industry | 0.232 | <.001 |
| Reflective Complex | 0.966 | <.001 |
| Intense Rebellious | 0.967 | <.001 |
| Upbeat Conventional | 0.900 | 0.030 |
| Energetic Rhythmic | 0.981 | 0.031 |

Note: A low p-value suggests a violation of the assumption of normality.

*Normality Test (Shapiro-Wilk) – High Quality Group vs. Low Quality Group*

| **Experimental Condition** | **W** | **p** |
|---|---|---|
| Age | 0.912 | <.001 |
| IML | 0.916 | <.001 |
| Musical Training | 0.731 | <.001 |
| Music Industry | 0.232 | 0.001 |
| Reflective Complex | 0.966 | <.001 |
| Intense Rebellious | 0.967 | <.001 |
| Upbeat Conventional | 0.900 | 0.032 |
| Energetic Rhythmic | 0.981 | 0.006 |

Note: A low p-value suggests a violation of the assumption of normality.

## 9.13 Appendix M

*Normality Test (Shapiro-Wilk) - Expectations*

| Experimental Condition | W | p |
|---|---|---|
| Article 1 & Article 2 | 0.994 | 0.768 |
| Article 1 & Control | 0.985 | 0.154 |
| Article 2 & Control | 0.985 | 0.131 |

Note: A low p-value suggests a violation of the assumption of normality.

*Homogeneity of Variances Test - Expectations*

| Experimental Condition | F | p |
|---|---|---|
| Article 1 & Article 2 | 0.033 | 0.856 |
| Article 1 & Control | 0.000 | 0.996 |
| Article 2 & Control | 0.031 | 0.861 |

Note: A low p-value suggests a violation of the assumption of equal variances.

*Normality Test (Shapiro-Wilk) – Vocal Evaluation (Control Groups)*

| Genre | W | p |
|---|---|---|
| Pop | 0.945 | 0.008 |
| Rock | 0.967 | 0.092 |

Note: A low p-value suggests a violation of the assumption of normality.

*Homogeneity of Variances Test - Vocal Evaluation (Control Groups)*

| Genre | F | p |
|---|---|---|
| Pop | 0.017 | 0.896 |
| Rock | 5.516 | 0.022 |

Note: A low p-value suggests a violation of the assumption of equal variances.

*Normality Test (Shapiro-Wilk) - Expectancy Violation*

| | W | p |
|---|---|---|
| Pop violation | 0.956 | 0.011 |
| Rock violation | 0.956 | 0.011 |

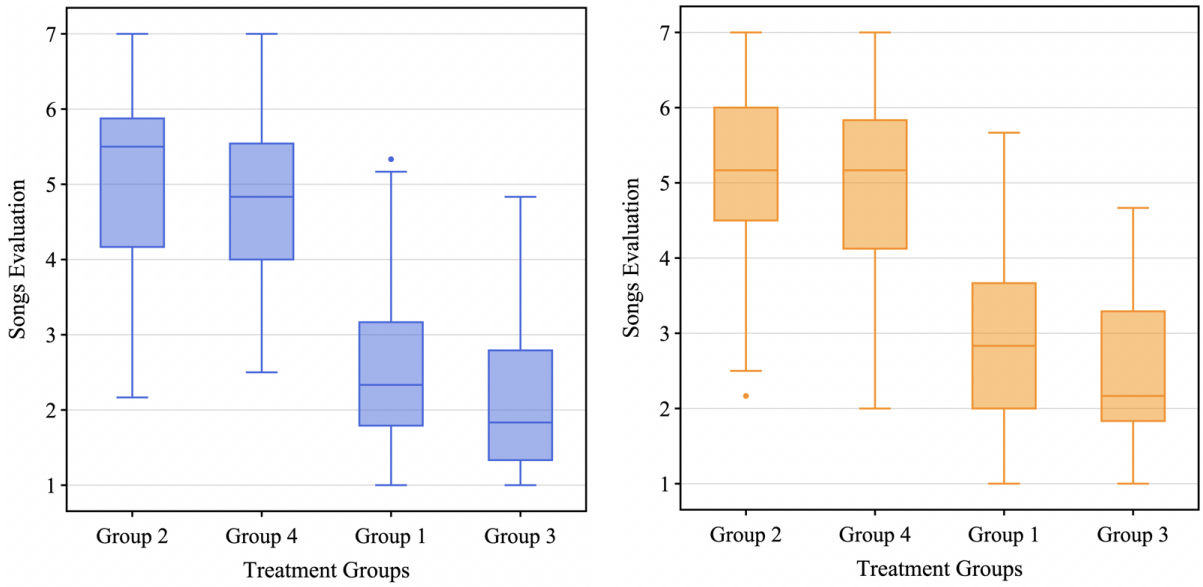
Note: A low p-value suggests a violation of the assumption of normality.

*Homogeneity of Variances Test (Levene's) - Expectancy Violation*

| | F | df | df2 | p |
|---|---|---|---|---|
| Pop violation | 10.559 | 1 | 74 | 0.002 |
| Rock violation | 0.997 | 1 | 74 | 0.321 |

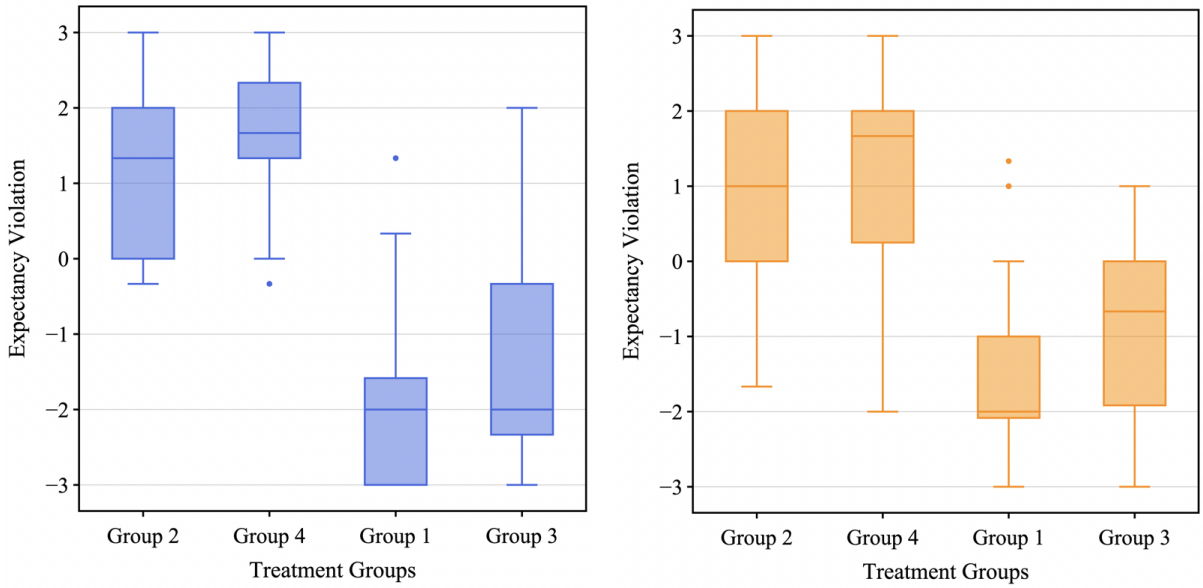
Note: A low p-value suggests a violation of the assumption of equal variances.

## 9.14 Appendix N




*(a) Pop* *(b) Rock*

*Box Plots: Songs Evaluation*




*(a) Pop* *(b) Rock*

*Box Plots: Expectancy Violation*

## 9.15 Appendix O

*Durbin-Watson Test for autocorrelation - Simple Regression Models*

| | Autocorrelation | DW Statistic | p |
|---|---|---|---|
| Model 1 | 0.642 | 0.713 | <.001 |
| Model 2 | 0.565 | 0.863 | <.001 |
| Model 3 | 0.653 | 0.692 | <.001 |
| Model 4 | 0.578 | 0.835 | <.001 |

Note: A low p-value suggests autocorrelation.

*Breusch-Pagan - Simple Regression Models*

| | Statistic | p |
|---|---|---|
| Model 1 | 0.356 | 0.551 |
| Model 2 | 0.007 | 0.932 |
| Model 3 | 1.69 | 0.193 |
| Model 4 | 2.36 | 0.125 |

Note: A low p-value suggests a violation of the assumption of homoscedasticity.

*Kolmogorov-Smirnov - Simple Regression Models*

| | W | p |
|---|---|---|
| Model 1 | 0.193 | <.001 |
| Model 2 | 0.188 | <.001 |
| Model 3 | 0.193 | <.001 |
| Model 4 | 0.182 | <.001 |

Note: A low p-value suggests a violation of the assumption of normality.

## 9.16 Appendix P

*Bootstrap Model 1*

| **Model 1** | $\beta$ | **BCa 95% Confidence Interval** | |
|---|---|---|---|
| | | **Lower** | **Upper** |
| (Constant) | 2.805*** | 2.040 | 3.509 |
| Attitudes_creative_AI | 0.224** | 0.043 | 0.414 |

Note: Bootstrap results are based on 2000 bootstrap samples.

*Bootstrap Model 2*

| **Model 2** | $\beta$ | **BCa 95% Confidence Interval** | |
|---|---|---|---|
| | | **Lower** | **Upper** |
| (Constant) | 2.927*** | 2.320 | 3.632 |
| Mean_score_attitudes_creative | 0.238** | 0.047 | 0.398 |

Note: Bootstrap results are based on 2000 bootstrap samples.

*Bootstrap Model 3*

| **Model 3** | $\beta$ | **BCa 95% Confidence Interval** | |
|---|---|---|---|
| | | **Lower** | **Upper** |
| (Constant) | 3.048*** | 2.390 | 3.671 |
| Mean_score_acceptance_role | 0.288** | 0.033 | 0.577 |

Note: Bootstrap results are based on 2000 bootstrap samples.

*Bootstrap Model 4*

| **Model 4** | $\beta$ | **BCa 95% Confidence Interval** | |
|---|---|---|---|
| | | **Lower** | **Upper** |
| (Constant) | 3.293*** | 2.686 | 3.933 |
| Mean_score_acceptance_role | 0.253** | 0.030 | 0.480 |

Note: Bootstrap results are based on 2000 bootstrap samples.

*Spearman's Rank Correlation Matrix*

| | ***pop_eval*** | ***rock_eval*** | ***attitudes_AI*** | ***accept_Singer*** |
|---|---|---|---|---|
| ***pop_eval*** | — | | | |
| ***rock_eval*** | 0.855*** | — | | |
| ***attitudes_AI*** | 0.170** | 0.197** | — | |
| ***accept_Singer*** | 0.142** | 0.147** | 0.364 | — |

Note: *$p < .05$, **$p < .01$, ***$p < .001$

## 9.17 Appendix Q

*Multiple Linear Regression Models Output with additional controls*

| **Variable** | **Model 5** | **Model 6** |
|---|---|---|
| Positive_expectations | 0.287 | 0.110 |
| | (0.240) | (0.256) |
| High_quality | 2.789*** | 2.362*** |
| | (0.229) | (0.244) |
| Positive_Expectations*High_quality | -0.065 | 0.055 |
| | (0.335) | (0.358) |
| Attitudes_creative_AI | 0.087 | 0.145 |
| | (0.074) | (0.110) |
| Acceptance_AI_singer | 0.136 | 0.063 |
| | (0.110) | (0.103) |
| IML | 0.026 | 0.014 |
| | (0.012) | (0.013) |
| Musical Training | 0.0144 | -0.089 |
| | (0.012) | (0.078) |
| Reflective Complex | -0.004 | 0.064 |
| | (0.073) | (0.104) |
| Intense Rebellious | -0.057 | 0.065 |
| | (0.098) | (0.080) |
| Upbeat Conventional | 0.040 | -0.080 |
| | (0.119) | (0.127) |
| Energetic Rhythmic | -0.082 | -0.011 |
| | (0.100) | (0.107) |
| Intercept | 1.627** | 2.285** |
| | (0.756) | (0.806) |
| Observations | 151 | 151 |
| $R^2$ | 0.832 | 0.783 |
| Adjusted $R^2$ | 0.692 | 0.612 |
| F Statistic | 25.9*** | 18.2 *** |

Notes: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

## 9.18 Appendix R

*Durbin-Watson Test for Autocorrelation - Multiple Regression Models*

| | Autocorrelation | DW Statistic | p |
|---|---|---|---|
| Model 7 | 0.036 | 1.91 | 0.510 |
| Model 8 | 0.064 | 1.81 | 0.170 |

Note: A low p-value suggests autocorrelation.

*Breusch-Pagan - Multiple Regression Models*

| | Statistic | p |
|---|---|---|
| Model 7 | 4.87 | 0.432 |
| Model 8 | 5.37 | 0.373 |

Note: A low p-value suggests a violation of the assumption of homoscedasticity.

*Variance Inflation Factors - Multiple Regression Models*

| | VIF | Tolerance |
|---|---|---|
| Attitudes_creative_AI | 1.46 | 0.684 |
| Acceptance_as_singer | 1.46 | 0.686 |
| Positive_expectations | 1.01 | 0.988 |
| High_quality | 1.02 | 0.985 |

Note: A VIF value above 10 suggests a violation of multi-collinearity.

*Normality Test (Shapiro-Wilk) - Multiple Regression Models*

| | W | p |
|---|---|---|
| Model 7 | 0.989 | 0.316 |
| Model 8 | 0.992 | 0.511 |

Note: A low p-value suggests a violation of the assumption of normality.

## 9.19 Appendix S

*Interaction test to verify the assumption of the homogeneity of regression slopes - Pop*

| **Source** | **SS** | **df** | **MS** | **F** | $\eta^2$ |
|---|---|---|---|---|---|
| Model | 2298.686 | 12 | 191.557 | 202.256*** | 0.946 |
| expectations*attitudes | 1.043 | 1 | 1.043 | 1.101 | 0.008 |
| quality*attitudes | 3.988 | 1 | 3.988 | 4.211 | 0.029 |
| expectations*_acceptance | 0.029 | 1 | 0.029 | 0.031 | 0.000 |
| quality*acceptance | 3.972 | 1 | 3.972 | 4.194 | 0.029 |
| expectations*quality*attitudes | 0.426 | 1 | 0.426 | 0.450 | 0.003 |
| expectations*quaity*acceptance | 2.378 | 1 | 2.378 | 2.510 | 0.018 |
| Error | 131.647 | 139 | 0.947 | | |
| Total | 2430.333 | 151 | | | |

Notes: a) R Squared = 0.946 (Adjusted R Squared = 0.941)
c) Type III Sum of Squares (SS)

*Interaction test to verify the assumption of the homogeneity of regression slopes - Rock*

| **Source** | **SS** | **df** | **MS** | **F** | $\eta^2$ |
|---|---|---|---|---|---|
| Model | 2424.890 | 12 | 202.074 | 175.492*** | 0.938 |
| expectations * attitudes | 1.291 | 1 | 1.291 | 1.121 | 0.008 |
| quality * attitudes | 1.728 | 1 | 1.728 | 1.501 | 0.011 |
| expectations * acceptance | 0.509 | 1 | 0.509 | 0.442 | 0.003 |
| quality * acceptance | 2.597 | 1 | 2.597 | 2.255 | 0.016 |
| expectations * quality * attitudes | 1.580 | 1 | 1.580 | 1.373 | 0.010 |
| expectations * quality * acceptance | 0.460 | 1 | 0.460 | 0.399 | 0.003 |
| Error | 160.054 | 139 | 1.151 | | |
| Total | 2584.944 | 151 | | | |

Notes: a) R Squared = 0.938 (Adjusted R Squared = 0.934)
c) Type III Sum of Squares (SS)

*Levene's Test of Equality of Error Variances*

| Model | F | df1 | df2 | Sig |
|---|---|---|---|---|
| ANCOVA POP | 0.762 | 3 | 147 | 0.517 |
| ANCOVA ROCK | 0.900 | 3 | 147 | 0.433 |

Note: Tests the null hypothesis that the error variance of the DV is equal across groups

*Normality Test (Shapiro-Wilk) - ANCOVA Experimental Conditions*

| | W | p |
|---|---|---|
| Condition 1 | 0.981 | 0.354 |
| Condition 2 | 0.971 | 0.179 |
| Condition 3 | 0.984 | 0.441 |
| Condition 4 | 0.948 | 0.004 |

Note: A low p-value suggests a violation of the assumption of normality.

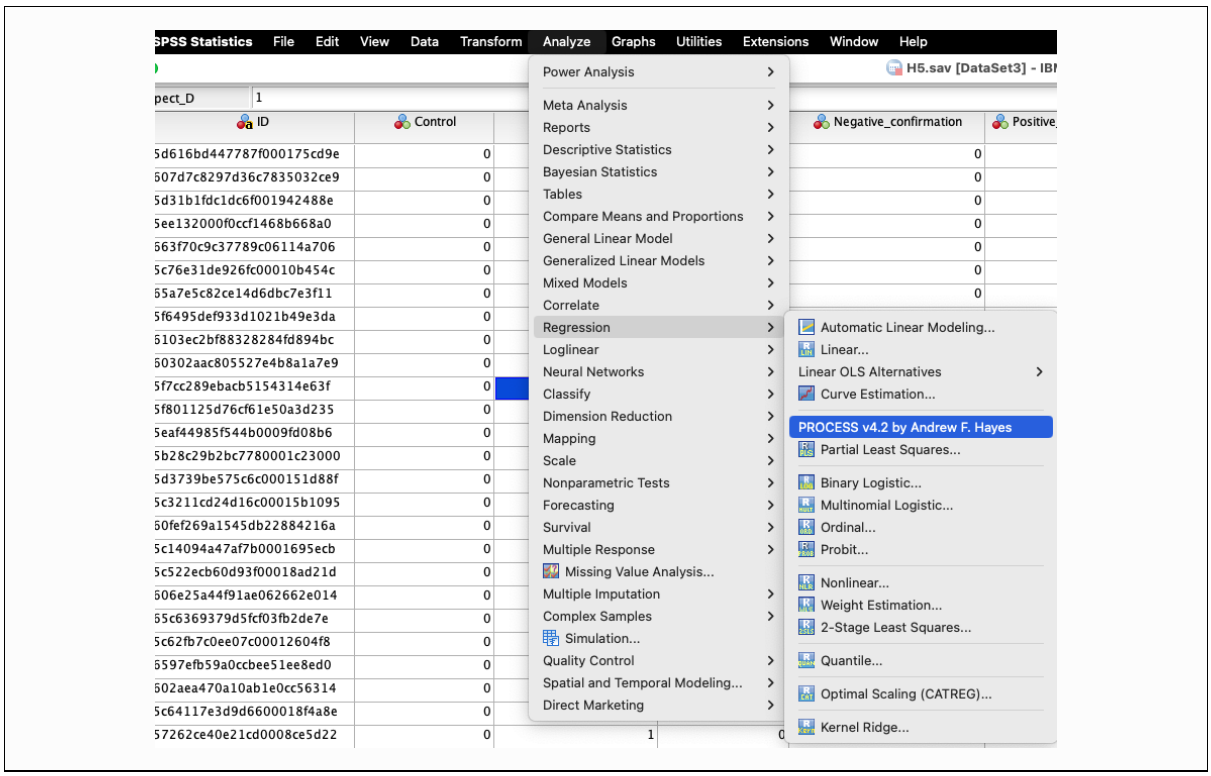
*Adjusted Means Pop*

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| Quality | Adj. Mean | Std. Error | Lower Bound | Upper Bound |
| Low Quality | 2.237[a] | 0.116 | 2.008 | 2.465 |
| High Quality | 4.998[a] | 0.114 | 4.773 | 5.223 |

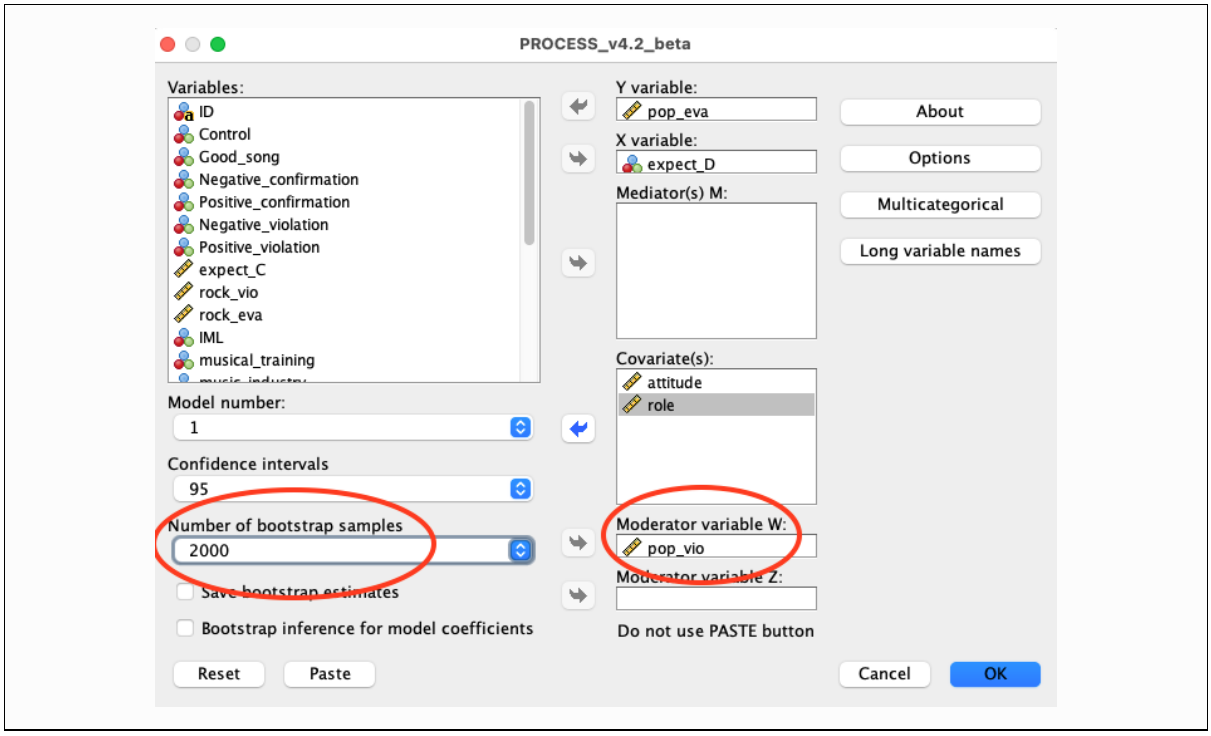
*Adjusted Means Rock*

| | | | 95% Confidence Interval | |
|---|---|---|---|---|
| Good_song | Mean | Std. Error | Lower Bound | Upper Bound |
| 0 | 2.574[a] | 0.125 | 2.326 | 2.821 |
| 1 | 5.006[a] | 0.123 | 4.762 | 5.249 |

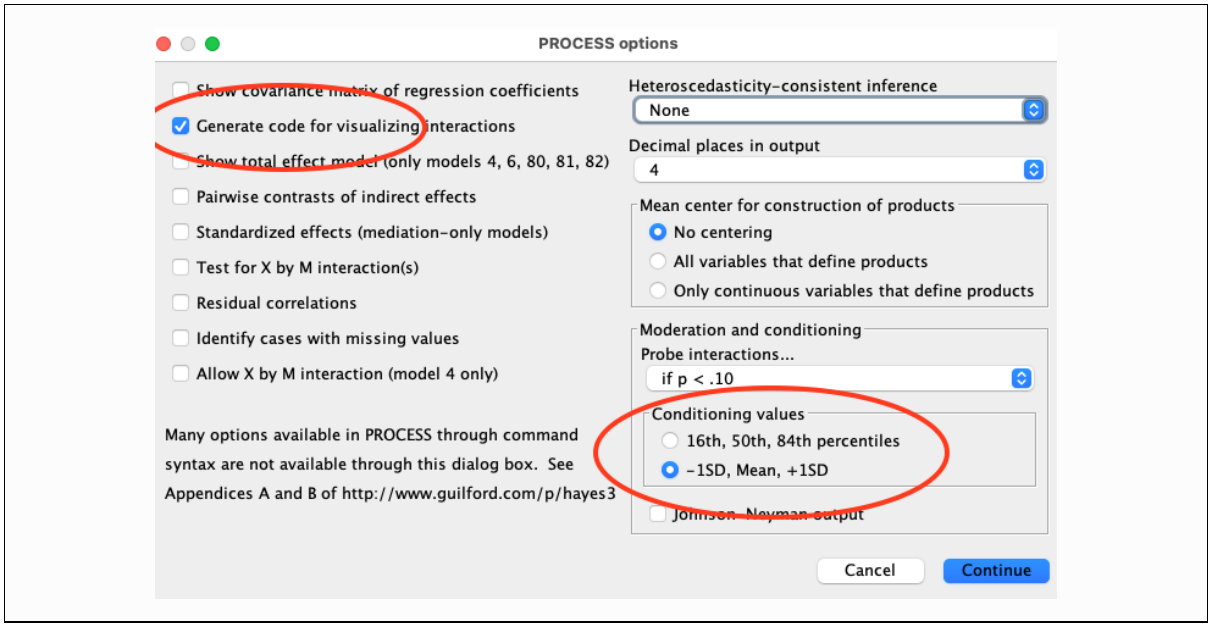
## 9.20 Appendix T




*Hayes's PROCESS macro: Step 1*




*Hayes's PROCESS macro: Step 2*




*Hayes's PROCESS macro: Step 3*