Affective Priming with Music: Cognitive, Psychoacoustic and Cultural Perspectives

ARMITAGE, JAMES,EDWARD

Affective Priming with Music: Cognitive, Psychoacoustic and Cultural Perspectives

James Edward Armitage

A Thesis presented for the degree of
Doctor of Philosophy

Durham University

Department of Music
Durham University
United Kingdom
January 2024

Abstract

This thesis explores the concept of affective priming with music: when a listener hears a musical sound just before being presented with a second stimulus, such as a word or image, the sound influences affective judgements about the second stimulus. For instance, in a word classification task in which words are classed as positive or negative, words are typically classified more quickly when preceded by a sound that has the same valence (e.g., a positive word is classified faster when preceded by a positive sound compared to a negative sound). This thesis considers how cognition, psychoacoustics and culture influence affective priming. Firstly, the thesis considers which dimension of emotion, valence or arousal, is transferred during priming. It then addresses how attention and affect interact in priming by considering the relationship between trait anxiety and affective priming. The results suggest that priming is the consequence of the interaction of top-down and bottom-up processes: primes engage attentional resources in a bottom-up process that conflicts with responses to the top-down word classification task. Chapter 8 considers whether stimulus features, namely harmonicity and roughness, influence priming. Harmonicity was not found to influence affective priming, whereas roughness was found to be a contributing factor to priming. Finally, the thesis compares results in priming and rating tasks for Western participants and Lithuanian Sutartinės singers. Whilst culture influences participants' ratings of musical stimuli, automatic evaluations, indexed by results of a priming task, are not influenced by culture. Combining the findings from Chapters 6 to 9, it is argued that affective priming is governed by the ability of the prime to demand allocation of attentional resources away from the target. Stimulus features which may contribute to the bottom-up process include acoustic roughness and the summation of musical features such as tempo, rhythm, dynamics and timbre.

Declaration

The work in this thesis is based on research carried out at the Department of Music, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

This PhD thesis is the result of support and kindness from many people. Firstly, I must thank Lesley for suggesting a mid-career switch to psychology. An honourable mention goes to Matt Watson for introducing me to embodied cognition, priming and reaction time methods – this all proved to be the inspiration for the PhD research. Vanessa's Parson provided an initial introduction to music cognition research. The last 6 years have been aided greatly by my fellow PhD students in the Music and Science Lab, with a special shout out to Scott, Matthias and Annaliese for their company, thought provoking conversation, lengthy discussions about R, and coffee. Kelly Jakubowski has been a great second supervisor, providing insightful assistance on design, methods and putting together the thesis as a whole. She is also an excellent violinist. Simone Tartisini provided invaluable technical support in preparation for a series of EEG experiments. Alas, Covid 19 intervened so these never reached fruition. Imre Lahdelma introduced me to the world of consonance and dissonance. I did not expect that his initial approach about using reaction time methods in consonance and dissonance research would lead to such a fruitful collaboration spanning several years, multiple publications and a partial change of direction for my PhD research. Finally, and above all, thanks to Tuomas Eerola whose willingness to take me on as a PhD student, unwavering positivity, thorough and rigorous approach, and generous spirit have made the last six years so rewarding.

Contents

List of Figures

List of Tables

Dedication

Dedicated to HHL

CHAPTER 1

Introduction

1.1 Overview

Music is capable of influencing judgements about words, images and people. This phenomenon has long been exploited by composers using devices such as word painting or *leitmotif*, and is employed in genres as diverse as pop songs, opera and film music. Alongside its creative applications, this phenomenon is used also in advertising. The ability of a stimulus (e.g., a sound) to influence our perception or cognition of a stimulus in another modality (e.g., an image or visually-presented language) is known as cross-modal priming. More specifically, where the influences are related to the affective or emotional properties (rather than the purely semantic), this phenomenon is known as cross-modal affective priming. This thesis considers this effect from the perspective of cognitive psychology and psychoacoustics, drawing predominantly on methods and theoretical frameworks from these disciplines, alongside consideration of the role of culture.

The thesis is based around six journal articles, five of which were accepted for publication in peer-reviewed journals at the time of submission, with one under review. The articles listed below therefore make up Chapters 4, 6, 7, 8 & 9 of the thesis. Chapter 10 (Armitage et al., 2023) has been subject to minor stylistic edits to accommodate the switch from numbered references to APA 7; otherwise the text of the chapters is unchanged from the published versions. The author contributions are outlined in Table 1.1. The articles are:

1. Eerola, T., Armitage, J., Lavan, N., & Knight, S. (2021). Online data collection in auditory perception and cognition research: Recruitment, testing, data quality and ethical considerations. *Auditory Perception & Cognition*, *4*(3-4), 251–280. https://doi.org/10.1080/25742442.2021.2007718

2. Armitage, J., & Eerola, T. (2020). Reaction time data in music cognition: Comparison of pilot data from lab, crowdsourced, and convenience web samples. *Frontiers in Psychology*, 2883. https://doi.org/10.3389/fpsyg.2019.02883

3. Armitage, J., & Eerola, T. (2022). Cross-modal transfer of valence or arousal from music to word targets in affective priming? *Auditory Perception & Cognition*, *5*(3-4), 192–210. https://doi.org/10.1080/25742442.2022.2087451

4. Armitage & Eerola (submitted) Anxiety induced biases in auditory affective

priming: a comparison of music and environmental sounds

5. Armitage, J., Lahdelma, I., & Eerola, T. (2021). Automatic responses to musical intervals: Contrasts in acoustic roughness predict affective priming in western listeners. *Journal of the Acoustical Society of America*, *150*(551). https://doi.org/https://doi.org/10.1121/10.0005623

6. Armitage, J., Lahdelma, I., Eerola, T., & Ambrazeviius, R. (2023). Culture influences conscious appraisal of, but not automatic aversion to, acoustically rough musical intervals. *PLoS ONE*, *18*. https://doi.org/https://doi.org/10.1371/journal.pone.0294645

Armitage and Eerola (2020) and Eerola et al. (2021) address methodological considerations; Armitage and Eerola (2022) and Armitage & Eerola (submitted) consider the cognitive mechanisms responsible for priming. Armitage et al. (2021) and Armitage et al. (2023) use priming as a vehicle to discuss current issues in consonance and dissonance research.

The thesis makes four contributions to the field. Firstly, the research has probed the cognitive mechanisms involved in affective priming in greater depth than previous music priming research. In particular, it has addressed the roles of attention and emotion in affective priming. Secondly, much like many priming studies in social cognition, it has used priming as a vehicle to probe attitudes towards stimuli and features of the stimuli. Third, it considers the role of culture in forming attitudes towards musical stimuli and explores differences between automatic and conscious evaluations of stimuli. Finally, alongside the theoretical contributions, the thesis has, from a methodological standpoint, added to the body of work that supports the use of online methods for data collection in auditory disciplines, something that became more into focus over the last three years. Moreover, to the author's knowledge, Armitage and Eerola (2022) was the first music cognition registered report to receive stage two acceptance.

The Covid 19 pandemic created a need for online data collection (and indeed, changed the course of this thesis somewhat). Fortuitously, Armitage and Eerola (2020) was already published as a proof of concept for online data collection in this context, and I developed a methodological research interest in online data collection alongside the main thrust of the thesis. Three months later, online data collection became a necessity: further consideration of the area led to Eerola et al. (2021). These two papers are presented in reverse chronological order in this paper to provide greater conceptual consistency in moving from the general to the specific.

The first part of the thesis introduces affective priming, establishes the existence of the effect and in Chapter 6 tests the reliability of online methods to detect congruency effects in affective priming. The thesis then goes on to consider some of the cognitive underpinnings of priming, in particular it asks which dimension of emotion, valence or arousal, is transferred from the prime to the target in affective priming. Further consideration of the cognitive mechanisms implicated in priming takes place in Chapter 8, which considers the role that attentional control plays in priming via group differences in trait anxiety. Chapters 9 and 10 use affective priming as a method to consider some questions in consonance and dissonance. In particular, Chapter 9 assesses the contribution of roughness and harmonicity to consonance and dissonance using affective priming as an indirect measure of attitudes to musical intervals. The final empirical chapter, Chapter 10, explores how cultural familiarity in a musical tradition that is rich in acoustic roughness, the Lithuanian Sutartinės style, influences automatic responses (i.e., affective priming) and conscious evaluative responses (i.e., likert scale ratings).

1.2 Affective Priming

Evaluations of affective stimuli are not frequently made on the basis of one modality alone. Such evaluations are often made in a broad affective context, for instance evaluations of advertising can be influenced by images or background music. An experimental paradigm exists that seeks to capture this empirically. Fazio et al. (1986) developed the paradigm as a measure of implicit attitudes to objects in an attempt to circumvent desirability effects. In the affective priming paradigm, two stimuli are presented in quick succession. Participants are typically asked to make a simple affective evaluation about the second stimulus (hereafter the *target*), such as categorising it as happy or sad, or as positive or negative. There is a broad consensus that the first stimulus (hereafter the *prime*) influences the response to the target. The usual dependent variables are reaction time and/or accuracy. The stimuli are typically affectively polarised, for instance a set of prime stimuli make consist of happy and sad faces, and the target set may consist of works associated with happiness and sadness. Broadly speaking, there is a consensus around the notion of *congruency effects*, i.e. a positive target is classified faster or more accurately when preceded by a positive prime than a negative prime; similarly, a negative target is classified more quickly and more accurately when preceded by a negative prime

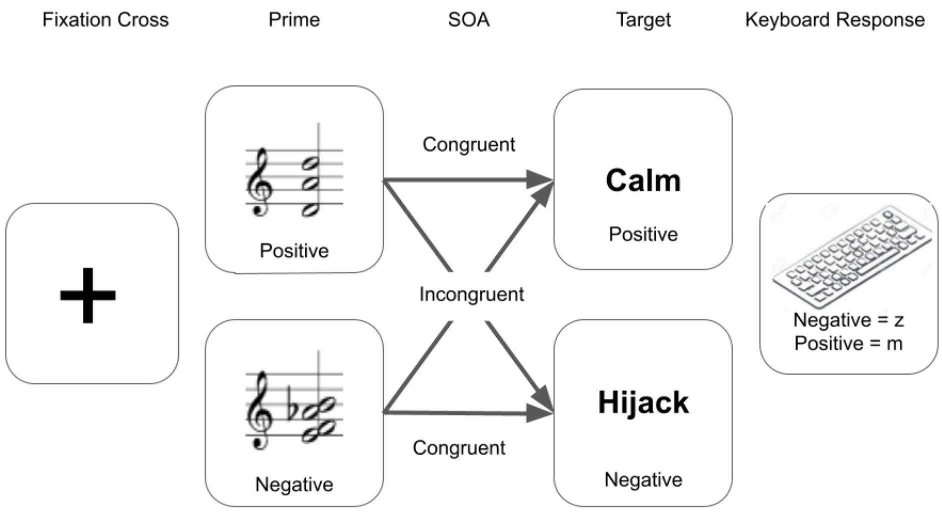


Figure 1.1: Schematic of affective priming experiment

than by a positive prime. For instance, if we consider a case where both the prime and target are words, the word *smile* would be categorised more quickly and more accurately when preceded by the word *sunshine* than when it is preceded by the word *painful.* This is illustrated in Figure 1.

In the affective priming literature, the term affective priming is used to refer to both the congruency effect and to the evaluative judgement task itself.

Affective priming is thought to depend on the automatic activation of attitudes (Fazio et al., 1986; Herring et al., 2013). However, despite the general consensus around the existence of priming effects, there is some disagreement around the exact mechanisms responsible for affective priming. Boakes (2010) and Klauer (1997) suggest two groups of explanations for affective priming: expectancy models and spreading activation models (Fazio et al., 1986). In essence, a key question in priming paradigms is whether the priming effect (usually operationalised as the difference in response times between congruent conditions and incongruent conditions) is

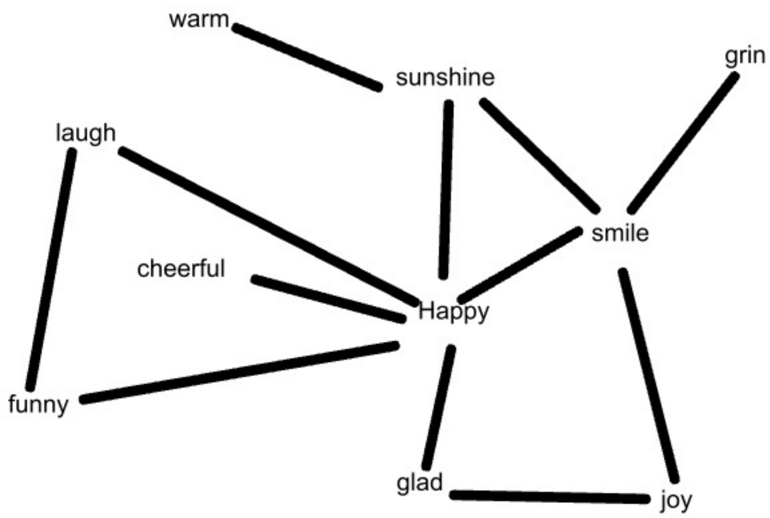


Figure 1.2: Example of a semantic network

a consequence of facilitation in congruent conditions (i.e., the spreading activation account), or of interference in incongruent conditions (i.e., the expectancy models). Spreading activation models explain affective priming in terms of activation in the semantic network. Most early accounts of affective priming argue in favour of spreading activation , namely that an affective prime activates a network of concepts closely associated with the prime. In particular, Fazio et al. (1986) argued that a prime activates the memory network, including any evaluative information. As a consequence, evaluative information about a congruent target is more readily accessible as it is already partially activated, resulting in faster response times to the evaluative classification task. Figure 1.2 shows a sample of a semantic network.

Variants of Fazio et al.'s explanation have been employed by several authors such as Bargh et al. (1988) and Hermans et al. (2002a, 2003). These authors suggest that primes are evaluated in broad terms such as positive and negative, and each evaluation has an individual node in the memory network. Furthermore, all positively valenced objects are interconnected, as are all negatively valenced objects via positively (and especially negatively) valenced nodes. Therefore, activation in one of these nodes makes other concepts linked to this node more accessible. A particular feature of this activation is that it takes place automatically (Hermans et al., 2001; Steinbeis & Koelsch, 2011). Figure 1.2 gives a diagrammatic representation of such a network. In music priming specifically, Steinbeis and Koelsch argue in favour of spreading activation accounts on the grounds that the very short time frame during which AP effects are evident suggests that it must be an automatic process.

On the other hand, interference accounts (Wentura, 1999) suggest that priming effects are a consequence of inhibition rather than facilitation. Zhang et al. 2006 assert that their behavioural findings support the idea of Stroop-like interference being responsible for inhibition in incongruent conditions. In the Stroop task, participants recognise the colour of a list of words. However, when the words themselves are names of the colours themselves that do not correspond to the colours in which they are printed. Reading speed is inhibited when the name of the colour and the colour of the ink do not match. Klauer and Musch (2003) argue that that an analogous mechanism provides a robust account of affective priming, but only in limited situations (i.e., in evaluation tasks rather than related paradigms such as the lexical decision task). Indeed, some of the theoretical explanations for facilitation, such as automatic vigilance, are difficult to reconcile in both positive and negative congruent conditions. An explanation based solely on interference rather than facilitation as a basis of priming effects automatically resolves some of the theoretical inconsistencies present in the other accounts. However, it should be noted that both interference and facilitation effects could be present in priming, i.e., a congruent activation is quicker than an unprimed activation, which is in turn quicker than an incongruent activation.

Alongside the cognitive explanations of affective priming, there is evidence that personality traits influence the extent to which participants exhibit affective priming. For instance, Blair et al. (2006) reported that psychopathy is linked with dampened affective priming effects (whereas other forms of priming were not significantly different between individuals high in psychopothy and typically developed controls.) Other traits which have been linked to priming are anxiety, empathy, neuroticism, gender, depression and openness to experience (Maier et al., 2003; Robinson & Tamir, 2005; Yamada & Decety, 2009). However, the results of these studies often suggest that the relationship between traits and affective priming is not always clear cut.

There is a small body of affective priming research in music cognition, which is reviewed in detail in Chapter 3. Music priming studies have mostly considered harmonic priming, e.g. using major vs minor chords or consonant vs dissonant chords as primes (for example Costa, 2013; Sollberger et al., 2003), or emotion perception in developmental disorders such as Alexithymia or Williams syndrome (Goerlich et al., 2011; Lense et al., 2013). In general, the predicted congruency effects have been present, with some exceptions depending on the modality of the target stimuli. Likewise, Williams syndrome and Alexithymia were associated with increased and decreased sensitivity to music priming respectively.

1.2.1 Thesis statement

The thesis proposed here is that, by considering the roles of psychoacoustics, culture and cognitive biases, we can better understand the mechanism responsible for cross-modal affective priming, that is to say how music influences judgments about stimuli in other modalities. More specifically, the aims of the thesis are threefold:

1. to establish the existence of cross-modal affective priming with music
2. to explore the cognitive and perceptual mechanisms responsible for affective priming
3. use affective priming as an implicit measure of attitudes to consonant and dissonant stimuli in order to assess the relative contributions of psychoacoustics and culture to consonance and dissonance at the level of automatic responses.

In addressing the aims above, the thesis has drawn on established methods from cognitive psychology and psychoacoustics. In particular the thesis uses reaction time methods that are discussed in Chapter 6.

1.3 Overview of the current research

Motivations for studying musically induced affective priming are threefold. Firstly, affective priming is a well established effect in a range of domains. However, there remains a degree of uncertainty around which cognitive and affective mechanisms are responsible for the effect. In particular, although priming effects have been probed in the case of music in a small number of studies, the mechanisms involved specifically in the case of music priming remain unexplored. This thesis will consider whether priming effects are a consequence of spreading activation or of stroop-like interference and whether the priming is a consequence of activity at the processing or the response stage. Moreover, the music priming literature has, to date, not focused on the underpinning mechanisms such as emotion, semantic activation or attention.

Secondly, as well as proving a source of research questions in itself, priming is used as a research method as an indirect measure of attitudes in social psychology and other areas, such as health psychology. By exploiting priming's utility in this respect, it is possible to use priming to probe unconscious or automatic responses to stimuli, giving insight to areas such as consonance and dissonance or emotion induction.

Finally, affective priming is important as a mechanism to explain compositional devices such as word painting, or in contexts where music can be used to influence judgements about stimuli in other domains, such as film music or advertising. Thus, an empirical and theoretical account of priming has applications in terms of explaining these musical phenomena.

The thesis will be organised as follows. Chapter 2 will present the theoretical framework, Cognitive Psychology, that is used throughout the research. Chapter 3 will present an historical overview of the broader research in affective priming from its origins in Fazio's seminal studies to the present day, with a specific focus on the competing explanations of affective priming, particularly with regard to spreading activation and Stroop-like interference accounts. Understanding the mechanisms affecting priming is central to understanding the cross-modal integration under investigation. The chapter will then focus more specifically on the literature on music and priming, beginning with Sollberger et al., 2003 and finishing with the most recent studies in musically-induced affective priming. The Chapter will present a meta-analytic synthesis of the studies most relevant to the current research.

Chapters 4 and 6 consider methodological issues, specifically the suitability of web-based methods for auditory research and specifically for affectve priming. Chapter 4 considers general issues in online data collection for auditory cognition and perception research. Chapter 5 introduces the reader to the main empirical chapters, providing an overview of the chapters that follow and detailing some common aspects of methodology and proposing a concrete model of priming based on the findings. Chapter 6 then discusses in detail the viability of collecting reaction time data using online methods, comparing data collected from a traditional lab sample with two online samples: a convenience sample collected via social media and a sample from a paid-for recruitment service.

The following four chapters make up the main empirical component of the thesis. In Chapter 7, the focus turns to the cognitive mechanisms of affective priming. This chapter uses a dimensional approach to emotion to address the question of which dimension of emotion, valence or arousal, is transferred in priming, or whether attention, via the feature-specific attention allocation (Spruyt et al., 2012), provides better predictions around transfer of emotion.

Chapter 8 considers whether individual differences in cognitive processing account for discrepancies in some previous priming literature. The chapter begins with a brief discussion of trait anxiety in terms of attentional bias to negative stimuli and introduced three theoretical models of anxiety: Beck's Information Processing account, Attentional Control Theory and the Vigilance-Avoidance model. The chapter goes on to introduce a pair of empirical studies to test hypotheses on the extent to which trait anxiety can modulate priming effects. Beyond any correlational relationship between anxiety and priming, the chapter uses a cognitivist approach to traits (i.e., a personality trait is considered in terms of the patterns of mental processing associated with it) to consider the relationship between anxiety, negative affect and attention, and how this is relevant to priming. The chapter tests which of three established motels of anxiety is most consistent with the priming effect. This chapter also assesses whether anxiety-related influences on priming are present for stimuli in different modalities (music and affective sounds).

The next two chapters use affective priming to probe properties of consonance and dissonance. This approach is parallel to that taken in social cognition, where priming is used as an implicit measure of attitudes. First, in Chapter 9, we use priming to explore automatic responses to consonant and dissonant musical intervals. In particular, it considers whether contrasts in acoustic roughness or harmonicity are responsible for automatic responses to intervals. Chapter 10 investigates the differential roles of culture and psychoacoustic factors in shaping automatic responses (i.e., responses captured by a priming paradigm) to acoustically rough intervals as opposed to conscious responses (via a rating scale) by comparing responses from Lithuanian participants (Sutartinės singers vs non-musician controls) and English-speaking participants (Western musicians vs controls).

In the final chapter, Chapter 11, the results of the research overall are discussed. A theoretical account of the music priming mechanism is proposed, taking into account stimulus features, cognitive mechanisms, particularly attention, and culture. The acccount is situated within the context of existing research on (music) priming. The scope and limitations of the thesis are discussed alongside proposals for future research.

Article	Author	Contribution
Armitage and Eerola (2020)	Armitage	First draft and revisions, critical comments, data collection and analysis, coding
	Eerola	data collection, critical comments, conceptualisation, funding acquisition
Eerola et al. (2021)	Eerola	First draft and revisions, critical comments, data collection and analysis, coding
	Armitage	First draft, data collection, critical comments, conceptualisation
	Knight	knowledge of linguistics papers, critical comments
	Lavan	knowledge of linguistics papers, critical comments
Armitage et al. (2021)	Armitage	First draft and revisions, critical comments, data collection and analysis, coding
	Lahdelma	Conceptualisation, critical comments, stimulus design
	Eerola	Data collection, critical comment, conceptualisation, stimulus design
Armitage and Eerola (2022)	Armitage	First draft and revisions, critical, data collection and analysis, coding
	Eerola	data collection, critical comments, conceptualisation
Armitage et al. (2023)	Armitage	First draft and revisions, critical comments, data collection and analysis, coding, funding acquisition
	Lahdelma	Critical comments, conceptualisation
	Eerola	Conceptualisation, critical comments
	Ambrazavicius	Map, recruitment, critical comments
Armitage et al. (2021)	Armitage	First draft and revisions, critical comments, data collection and analysis, coding
	Eerola	Conceptualisation, critical comments

Table 1.1: Author contributions for papers

CHAPTER 2

Cognitive Psychology in Music Research

Abstract

This chapter will present the methodological frameworks that are appropriate for tackling the research questions outlined in Chapter 1. Although growing in popularity and influence, widespread use of experimental approaches to musicology is still a relatively young phenomonen musicological questions are still typically tackled within methodological approaches drawn from the humanities. However, the publications within the present thesis owe more to cognitive psychology than to musicology in the humanities tradition. This chapter begins by motivating the need for an empirical approach to musicology and outlining the scientific paradigm - the hypothetico-deductive method that will be used throughout the research. It will

then introduce the main disciplinary and methodological areas that are used in this thesis: cognitive psychology and psychoacoustics and consider how psychoacoustic features before considering how culture can influence cognition. The chapter concludes by outlining the implications for the present research.

2.1 Cognitive Psychology Methods in Music

Music cognition sits at the interface of musicology, music theory and cognitive psychology. Before addressing the research questions outlined in Chapter 1, it is necessary to establish which set of approaches is best equipped to fulfil that purpose. This leads to a number of key questions. For instance, what does music cognition bring to the study of music that established disciplines such as historical musicology or analysis do not? How does music cognition add to our understanding of music and does studying music help psychologists and cognitive scientists understand human mental processing more generally? In the context of the present research, it is important to note that affective priming is largely studied within cognitive psychology and allied domains, such as linguistics and cognitive neuroscience.

2.1.1 Cognitive psychology

Cognitive psychology (Neisser, 2014) was, to some extent, born in contrast to the behavioural school associated with Skinner. Cognitive psychology as a distinct subdiscipline of psychology is concerned with internal mental processes. Cognitive psychology is considered both a branch of psychology the study of the human mind and behaviour and of cognitive science more broadly. The so-called Cognitive Revolution encompassed aspects of psychology alongside philosophy, linguistics, computer science and anthropology (see e.g., Miller, 2003). Barsalou (2014) describes the internal mental processes (cognitive constructs) thus:

> The internal constructs of cognitive psychology tend to be of a very specific type: almost always they describe information processing, describing how the brain processes information. Some cognitive constructs correspond to mechanisms in the brain that pick up information from the environment; others to mechanisms that store information in memory; and still others to mechanisms that retrieve information from memory, transform information in memory, and send information back into the environmentnearly all cognitive constructs describe information processing mechanisms (Barsalou, 2014, p.9)

The cognitive approach stands in contrast with *behaviourism* (Watson, 1913), which had been the dominant model in psychology for much of the mid-twentieth century. Behaviourism was concerned with human behaviour as a response to conditioning: the underlying processes were not considered in depth. The processes and mechanisms that cognitive psychology deals with include memory, language, attention, vision, emotion, and attention. It should be emphasised that cognitive psychology typically is concerned with the processes and not the neural structures that underpin them (this is the domain of cognitive neuroscience). Some authors consider the division between cognitive psychology and (cognitive) neuroscience as being akin to the relationship between software and hardware in computing (for instance, Boone & Piccinini, 2016). Furthermore, the outward behaviour as distinct from the internal mental process associated with it does not fall within the domain of cognitive psychology beyond the extent to which observation of the outward behaviour informs study of the internal mental process. Next we consider some of the mental processes that are central to the present thesis. A concept central to the present thesis is attention. In general, attention can be split into two types. Firstly, top-down attentional processes are defined as processes that are strategic, i.e., an observer is volitionally deploying cognitive resources to a particular task or stimulus and not others. The second type of attention is bottom-up attention. Bottom-up processes occur when some feature of a stimulus demands attentional resources whether or not this fits with the observers plans. (For an overview of top-down and bottoom-up processes see Katsuki & Constantinidis, 2014; Pinto et al., 2013) The need for attention is a consequence of the brains limited capacity to process information. Some authors, e.g. (e.g. Kahneman, 1973) conceptualise attention as a finite resource. Another model of attention is the perceptual load theory, which splits attention into perceptual and cognitive components. Under this model, perceptual attention is an individuals ability to attend to or filter out a particular stimulus; cognitive attention is the individuals ability to carry out the internal mental process relating to the stimulus. Relatively speaking, attention has not formed a central aspect of music cognition research. However, there are notable exceptions, such as Bigand et al. (2000) and Tervaniemi et al. (2009).

A second concept drawn from cognitive psychology is *emotion* (although there is a degree of tension as to whether affective science forms a separate discipline, e.g., Davidson et al., 2009). Emotion is typically operationalised using one of two models to describe the varieties of emotional states. The first is the basic emotion model (Ekman, 1992). In Ekmans highly influential conceptualisation of emotion, there are six fundamental or basic emotions, namely happiness, sadness, anger, surprise, fright and disgust. More complex emotions are made up of combinations of the basic emotions. Other conceptualisations of basic emotion exist, for instance Hutto et al. (2018). The second set of models of emotion that are employed by emotion researchers are the so-called dimensional models of emotion. In these models, emotions are not described by semantic labels as in the basic emotion models. Rather, they are mapped to a two or three dimensional space. One example of such a model is the valence-arousal (or circumplex) model of emotion (Russell, 1980). This model represents emotions in two-dimensional valence-arousal space, such that happiness would be described as a positive-valence high-arousal emotion, anger a negative valence high arousal emotion and sadness a negative valence - low arousal emotion. The relationship between music and emotion has received much attention in recent years. Fundamental mechanisms of emotion induction, strong emotional reactions, recognition, emotion regulation, memory and emotion and conveyance of emotion in performance have all been the focus of a significant amount of research activity

Many authors define cognitive psychology as an approach, i.e., as well as being a body of particular psychological phenomona (internal mental processes) that are to be studied, cognitive psychology is a way of explaining human human behaviour more broadly in terms of the internal mental processes. Indeed, Chapter 7, which uses a trait measure of anxiety, which is often situated with the individual differences approach, still situates the traits approach within cognitive psychology by emphasising the relationship between trait anxiety and the pattern of information processing associated with it. Defining a potentially static construct such as a trait in terms of the internal mental processes with which the trait is associated is known as the *cognitivist* approach. Likewise, several questions in psychology – and indeed music psychology specifically – can be approached from the cognitive perspective.

As it is rooted in the study of observable behaviour, and is concerned with developing models of mental process through the study of observable behaviours, it is crucial to consider cognitive psychology as an empirical discipline. That is a discipline in which theories and models are tested by experimental means. Moreover, it is a fundamental assumption of cognitive psychology that cognition can be studied via the scientific method (albeit an assumption that has been the subject of considerable debate; see, e.g., Haig, 2014). The dominant paradigm within the scientific method is the hypothetico-deductive method. Indeed, the hypothetico-deductive method is frequently cast as *the* scientific method. This method is frequently employed, often tacitly, by scholars in cognitive psychology, and is underpinned philosophically in the work of authors such as Popper. Niiniluoto (2011) summarises the hypothetico-deductive method as follows:

> The hypothetico-deductive (HD) method of science allows scientists to freely invent hypothetical theories to explain observed data, but requires that such hypotheses are indirectly tested by their empirical consequences. (Niiniluoto, 2011, p.338)

Thus, when we consider cognitive psychology as a scientific discipline, we typically mean, one in which the hypothetico-deductive method is followed: hypotheses are formed on the basis of theory, experiments are designed to test the hypothesis, data is collected and analysed, and conclusions related to the hypotheses are formed on the basis of the analysis of the data. Should the data align with the hypothesis, this is considered as confirmatory evidence of the hypothesis, whereas if the data does not support the hypothesis, the experiment is considered to have disproved the hypothesis. Despite its broad base of support and use within the scientific community, there have been criticisms of the method.

Cognitive psychology is concerned with all mental information processing including the processing of sound. Neisser (2014) describes auditory cognition as being concerned with transformation of the fluctuating pressure-pattern at the ear into the sounds and the speech and the music that we hear (p.4). On this basis, we consider auditory cognition as encompassing the internal mental processes that take place once a sound has passed through the auditory system and been passed to the auditory nerve. However, in order to understand fully these mechanisms, we need also to turn to a domain that allows us to understand the nature of the stimuli to be processed and how their passage through the auditory system influences cognition. To this end, we next discuss psychoacoustics.

2.1.2 Psychoacoustics

In its most encompassing definition, Acoustics is defined as the scientific study of mechanical waves, including of most relevance to the present thesis sound waves. Its subdiscipline psychoacoustics sits at the intersection of acoustics and cognitive psychology and is concerned with studying the properties of sound in terms of their mechanical features and how these sounds propagate through the (human) auditory system. Psychoacoustics thus considers sound waves not only from a mathematical point of view but also as sensory and perceptual events. Psychoacoustic methods allow scholars to quantify features of auditory stimuli in describing their pitch, loudness and timbre and to model the physiological impact of these features on the inner ear. Moreover, psychoacoustics provides the means to interpret these measures in the context of the auditory system. Although scholars have sought to explain how humans perceive sound, and in particular consonance and dissonance, since Classical Antiquity, modern psychoacoustics has its genesis in the work of von Helmholtz, and in particular his 1863 work *On the Sensation of Tone as a Physiological Basis for the Theory of Music*, (Helmholtz, 1885) remains an influential text in the field. Helmholtz main contribution was his spectral approach the notion that the inner ear can decompose complex tone into its constituent sinusoidal partials, suggesting that the inner ear carries out a Fourier analysis of a sound signal. Indeed, this remains the basis of current theories of pitch perception.

The present research uses the psychoacoustic measurements of roughness and harmonicity. Roughness is a percept characterised by rapid amplitude modulation of a sound. Roughness occurs in the liminal space where two sinusoidal waves are too close to be perceived as separate, but too far apart to be perceived as a unison. Subjectively, roughness is usually perceived as unpleasant. There are several models of roughness, (e.g., Hutchinson & Knopoff, 1978; Wang et al., 2013). Whilst frequently used as a measure of nuisance noise in contexts such as car engines, railways or urban environments. In addition, in a musical context, roughness has been applied to the study of timbre and of consonance and dissonance. Another important construct in psychoacoustics is harmonicity. Harmonicity is the phenomenon whereby the auditory system groups together frequency components that are thought to have a common fundamental frequency (hereafter F_0). Most models of harmonicity make the assumption that two sounds with a common fundamental frequency, F_0, originate from a common source; where waveforms suggest two different F_0s, this is perceived as two separate pitches. A high degree of harmonicity is typically associated with concepts such as pleasantness and consonance. Conversely, lower degrees of harmonicity are associated with dissonance or unpleasantness. Both roughness and harmonicity have been central to modern discussions of consonance and dissonance. However, the relative contribution of each (and other factors such as culture) is the subject of much scholarly debate. As is the case with roughness, there exist several models of harmonicity (Harrison & Pearce, 2020). However, models of both roughness and harmonicity have had limited success in predicting empirical results. It is important to note that most models of roughness and harmonicity are based on models of the mechanical properties of the sound waves. When the mechanical properties of the waves are understood, it is possible then to consider the effect these waves have on the human auditory system and in turn on cognition.

2.1.3 Empirical approaches to music

Empirical approaches to music draw on methodologies from a wide range of established (scientific) disciplines: psychology, physiology, computer science, and anthropology amongst others. Widespread implementations of empirical approaches to music are relatively modern, though there are historical antecedents in the work of scholars such as Stumpf (1898) and Helmholtz (1885). Indeed, scientific approaches to music have earlier antecedents, for instance as early as the late 17th Century. This is captured neatly by Cohen (2013): ...there is another, scientific, approach to music, which takes as its starting point, not compositional techniques, but musical sound. In this approach, music is not looked at in terms of aesthetics, but in terms of mathematics, physics and physiology. Since compositional techniques, in the last analysis, come down to highly differentiated applications of musical sound, it may be legitimately asked whether ultimately a complete reduction of the musical experience to physical and physiological mechanisms might be achieved. (Cohen, 1984, pXI)

What distinguishes scientific approaches from those found in the humanities tradition of musicology is not necessarily the research questions – questions such as *why are we moved by sad music?*, *what constitutes dissonance?*, *why does a particular chord make sense in a progression but another does not?*, *what would a 19th century audience have made of this?* can all be approached from a variety of methodological standpoints from within either traditional humanities-based musicology and cognitive psychology. The difference is in the implementation of the methods and in particular methodologies used to approach the problems. Although it should be noted that certain questions that are well-posed in music cognition may not be worthwhile in the humanities tradition. For instance, a question such as *"is melodic expectation a product of schematic or statistical learning processes?"* provides fertile ground for study in cognitive psychology but might prove intractable to humanities-based research methods. Similarly, there are research questions that can be addressed effectively using methods from traditional analysis or historical musicology that empirical musicology would struggle to shed any light on. Overall, music has many features that make it particularly well suited to be subject to methods from cognitive psychology. Firstly, despite a large degree of variation in the nature of music making between and within cultures, music in some form or other appears to be present in all cultures (Blacking, 1995; Mehr et al., 2021), and indeed there are indications that music has been present for as long as homo sapiens has been around. Thus the mental processes associated with music are almost certainly to some extent universal even if music and the practice of music making vary from culture to culture. Moreover, many of Barsalous 2014 major cognitive constructs are necessary to produce or consume music: attention, emotion, memory, language, perception. Finally, music is – in many contexts – a cross-modal phenomenon. Both vision and movement contribute to how we mentally process music. As a result, approaches to the study of music that employ methods drawn from cognitive psychology are particularly fruitful. Despite the growth in research activity in music psychology in recent years, there is nevertheless a need to consider further how information from different modalities integrated into a listener's experience of music.

Music is, of course, an activity that is deeply rooted in culture. When considering the cognition of music, we also need, therefore, to consider the extent to which culture can affect cognition. It is self-evident that music exists as a product of human activity and that music varies over both location and time. In particular, the question of whether attitudes – and particularly evaluations of musical stimuli – are shaped by culture is crucial to the present thesis, and this question is addressed in Chapter 9. There seems to be a strong accord that culture plays an important role in shaping cognition. Cultural psychologists talk of field dependence, defined by Ji and Yap as the degree to which the perception of an object is affected by contextual factors surrounding the object (Ji & Yap, 2016). In the context of music, field dependence can be thought of both immediately, i.e., how does a specific musical feature relating to melody, harmony, rhythm, or timbre fit in the immediate musical context it is being presented (often studied in the context of musical expectation), or on a broader level – that is how does a musical feature (or piece of music) fit into the musical culture at large. Deličge and Sloboda (2004) provide a neat summary:

> The research literature, therefore, leads to the conclusion that human beings pick up quite high-level implicit (or tacit) knowledge about some major structural features of the music of their culture. (Deličge & Sloboda, 2004, p.208).

Familiarity is a contributing factor to how culture shapes attitudes. However, some authors have argued that there is a confound in the relationship between familiarity and attitude relating particularly to automatic evaluations: that paradigms such as the implicit association task access familiarity rather than attitudes. An additional layer beyond familiarity is expertise. Indeed, one of the hypotheses tested in Chapter 9 is the extent to which expertise can influence attitudes to musical stimuli and whether or not attitudes that psychoacoustic properties that would typically result in negative evaluations can be changed under the influence of culture, operationalised here as expertise and familiarity.

2.1.4 Implications for the present thesis

The present thesis seeks to explore the concept of affective priming, i.e., the phenomenon in which evaluations of a stimulus are influenced by the presence of an immediately preceding stimulus. It draws on use methods from cognitive psychology to explore the internal mechanisms responsible for affective priming and to use methods from psychoacoustics to probe the relationship between stimulus features, particularly measures such as roughness and harmonicity, and those cognitive mechanisms. Previous research has implicated attention, semantic memory and emotion in affective priming and consequently the topic falls squarely within the domain of cognitive psychology. More generally, the paradigm used in the present thesis is an example of a cross-modal phenomenon, i.e., where perception or sensation of a stimulus in one sensory modality influences perception, sensation or processing of a stimulus in another modality. Affective priming, and reaction time methods more generally, fall within the tradition of mental chronometry - using measurements of reaction time to shed light on mental processes. Mental chronometry is a long-established method in cognitive psychology, employed rigorously since the mid-19th Century beginning with Francis Galton. Mental chronometry assumes that mental processes take place in real time. Posner outlines the role of mental chronometry in cognitive psychology and other disciplines below:

> Mental chronometry can be defined as the study of the time course of information processing in the human nervous system. The stress on information processing and humans indicates that mental chronometry seeks systematic experiences that relate to them. As such, mental chronometry serves as a means of relating the differing viewpoints of phenomenology, physiology, and performance in so far as the three languages can be applied to common topics close to the boundaries of their application (Posner, 1978, p.8)

Indeed, mental chronometry has been used, albeit relatively infrequently, in music cognition to study harmonic relatedness (Bharucha & Stoeckig, 1986, 1987), in affective priming studies (e.g., Sollberger et al., 2003; Steinbeis and Koelsch, 2008; see Chapter 3 for a more comprehensive review), in hook identification (Burgoyne et al., 2013), and to probe harmonic expectation violation (e.g., Janata, 1995) among other topics. Thus, although much of the research in the present thesis draws on a methodological tradition from outside music cognition, reaction time (and more specifically affective priming) methods are nevertheless well-established vehicles for studying musical phenomena.

Chapter 6 considers attention and emotion and in particular whether focusing attention on one particular aspect – or dimension – of emotion can manipulate the results of an affective priming task. Attention is considered from both top-down and bottom-up perspectives.

Chapter 7 uses the cognitivist approach to personality in order to consider the role of a fundamental concept in cognitive psychology, attention, in affective priming. Chapters 8 and 9 incorporate methods from psychoacoustics to study properties of the prime stimuli in terms of there harmonicity and in particular roughness. Chapter 9 then contrasts the influence of roughness and culture, using a combination of affective priming and likert ratings.

2.1.5 Conclusions

Psychological, and especially cognitive, approaches to the study of musical phenomena have grown in popularity in recent years, i.e., questions of how music is processed internally. The present thesis will address the topic of cross modal affective priming using musical prime stimuli – i.e., the questions of whether, when and how music can influence evaluative judgements about non-musical stimuli. The topic will be addressed from the standpoint of cognitive psychology to allow for an examination of the internal mental processes that contribute to affective priming. In particular, the implication of memory, attention and emotion in affective priming makes cognitive psychology the most appropriate framework to consider this topic. Additionally, methods from psychoacoustics (considered as a subdiscipline of both cognitive psychology and acoustics) will be used to consider which stimulus features are associated with priming. Finally, we use affective priming in the social cognition tradition as an indirect measure of attitudes to the stimuli and will consider whether attitudes shaped by culture (expertise and learning) influence responses to stimuli which are predicted to be associated with negative attitudes by psychoacoustics. Throughout this thesis, the hypothetico-deductive method will be the governing scientific paradigm.

CHAPTER 3

Affective Priming: an historical overview & meta-analytic review

Abstract

This chapter begins with a historical overview of the development of the affective priming paradigm from the seminal affective priming studies by Fazio in the mid-1980s (Fazio et al., 1986) until the present day. Drawing on literature from a range of psychological domains, the chapter discusses the existence of congruency effects, the cognitive mechanisms (i.e., the question of "how does priming work?") and the applications of priming as an automatic measure of attitudes. After a historical overview of general priming research, the literature review then focusses more specifically on priming in music and auditory domains. The chapter concludes with

a meta-analysis of the studies most relevant to the present thesis. Moderating factors such as stimulus onset asynchrony, prime type (music vs consonant/dissonant chords vs major/minor chords), target type (words vs images) and number of trials were considered. Stimulus onset asynchrony and number of readings were significant predictors of effect size. At low stimulus onset asynchrony (<500 ms), effect sizes were greater than at high stimulus onset asynchrony; larger numbers of readings were associated with smaller effect sizes. There was a significant interaction of prime type and target type.

3.1 Introduction

Affective priming is a well-established experimental paradigm that has been used as a measure of implicit attitudes in both cognitive and social psychology for nearly 40 years since it was first studied by Fazio and colleagues in 1986 (Fazio et al., 1986). The literature review begins with a historical overview of the priming literature from its origins to the present day. This first part of the review spans literature from several domains in psychology, such as social cognition, memory and psycholinguistics. The review outlines research which considers the conditions necessary for priming effects to be detected and discusses the different theoretical models that explain priming, particularly focusing on spreading activation accounts as compared to interference accounts. The chapter then considers how priming can be used as an implicit measure of attitudes. Once the main strands of the broader literature have been detailed, the review turns to affective priming in music psychology and other auditory domains. The literature review concludes with a meta-analysis of the studies that are most relevant to the present thesis. This chapter focusses on behavioural (i.e., reaction time and accuracy rate) data, although incorporates EEG data where appropriate.

Affective priming has its origins in semantic priming studies that became popular in the late 1960s and 1970s. These studies were concerned with testing the *spreading activation* model of information retrieval (Quillian, 1967). Quillian's model conceptualises words as nodes in a network with associations between the words as edges. Quillian asserts that in retrieving information from a node, other, most nearly connected, nodes are also activated. One of the predictions of the spreading activation model is *semantic priming*. Semantic priming tasks, such as the lexical decision task (LDT; Meyer & Schvaneveld, 1971), typically involve priming a non-affective category associated with the target. In the LDT, participants are presented with strings of letters which either meaningful (i.e., they make up a word), or they are not meaningful (i.e., they make up a non-word). Participants identify a string of letters as being a word (target) or non-word. In the semantic priming variant of the LDT, the words/non-words are preceded by a prime stimulus. The primes are typically words that are orthogonally manipulated such that they are either semantically related or unrelated to the target. Responses are faster and more accurate when the prime and the target are semantically related. For example, the word *fork* would be classified more quickly and accurately when preceded by the word *spoon* as compared to the word *garage*. Figure 3.1 outlines the procedure for such a task.

Fazio introduced the affective priming paradigm in 1986 as a measure of automatic activation of attitudes (Fazio et al., 1986). Early affective priming studies

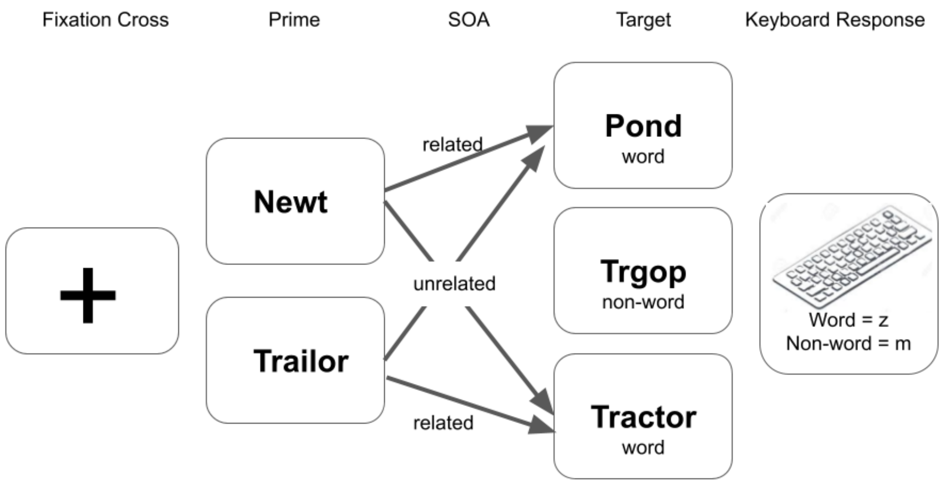


Figure 3.1: Outline procedure for lexical decision task

explained the phenomenon in terms of the spreading activation account, that is, a stimulus (typically a word in the early studies) would activate affectively related concepts in the semantic memory. Early accounts were based on a number of assumptions (see, e.g., Fazio, 1986). Firstly, that humans evaluate stimuli automatically in simple polarised terms, e.g. good-bad. The purpose of such evaluations is to allow humans to operate in an environment containing potential threats or positive opportunities (Klauer, 1997). It is assumed that these evaluations take place at an early stage of emotion processing. In essence, these automatic evaluations correspond to approach-avoidance tendencies. Next, early accounts assume that responses to stimuli are automatic or reflex-like. Finally, early accounts assumed that automatic evaluations are governed by spreading activation through affective-semantic networks. The review considers these early accounts and more recent developments which have considered alternative models of the cognitive processes underpinning priming.

Following Fazio's initial exploration of affective priming, the paradigm became more widespread, and researchers probed both priming itself alongside using affective priming as a method to explore concepts in other domains such as social psychology, mental health, and linguistics by considering the circumstances under which affective priming occurred and by exploring the mechanisms responsible for affective priming. An influential body of work was contributed by Hermans, Spruyt, de Houwer and colleagues (for a reeview see De Houwer et al., 2009). They probed automatic activations (Spruyt & Tibboel, 2015), semantic relatedness (Houwer et al., 2002), strategic influences on priming (Spruyt et al., 2017), the time course of priming (Hermans et al., 2001), expectency and learning (Hermans et al., 2002b), feature-specific attention allocation (Spruyt et al., 2009, 2012), and the relationship between trait anxiety and priming (Hermans et al., 2003). Voss et al. (2013) employed a novel analytic strategy (drift diffusion modelling) to prise apart the differences between affective priming and semantic priming, arguing that semantic priming was driven by spreading activation whereas affective priming is driven by response competition (see below for a discussion of the two accounts).

As well as research that has explored the cognitive mechanisms responsible for priming, there is also a more applied body of research that has used priming as a research method in order to probe attitudes in a variety of settings. Initially this was extensively used in social cognition to study stereotyping, for instance, racial (Livingston & Brewer, 2002) or gender (van Breen et al., 2018) stereotyping before being applied in a wider range of contexts, some of which are outlined below.

In addition to revealing attitudes to certain stimuli, affective priming has been used to study individual or group differences in automatic processing that can be accounted for by personality traits, e.g., in anxiety (Li et al., 2007; Maier et al., 2003), psychopathy (Blair et al., 2006; Meier et al., 2007), empathy (Yamada & Decety, 2009), and neuroticism (Robinson et al., 2007). Whilst affective priming does not necessarily offer a measure of a specific trait, priming research has been successful in capturing cognitive processes that are present in the case of some populations but not others.

There have also been applications of affective priming in bilingualism, (Degner et al., 2012; Tenderini et al., 2022). Tenderini et al. considered the relationship between age of second language (L2) acquisition and affective processing in L2 via an affective priming paradigm. Degner et al. used affective priming to consider the relationship between affective processing and frequency of L2 use. Most recently, affective priming has been applied in fields as diverse as health behaviours during the Covid-19 pandemic (Moro & Steeves, 2021, 2023; Scerrati et al., 2022), attitudes towards diet (Legget et al., 2022; Tzavella et al., 2020), advertising (De Luca & Botelho, 2020; Li, 2021), and neural correlates of cross-modal processing (Tiihonen et al., 2024).

3.2 Mechanisms of affective priming

3.2.1 Spreading Activation Models

Spreading activation models can be subdivided into affective evaluative activation and semantic evaluative activation. Most early accounts of affective priming argue in favour of spreading activation (Fazio et al., 1986; Hermans et al., 2001; Sollberger et al., 2003), namely that an affective prime activates a network of concepts closely associated with the prime. In particular, Fazio et al. argued that a prime activates the memory network, including any affective evaluative information associated with the prime. As a consequence, evaluative information about a target that is congruent to the prime is more readily accessible than an incongruent target, as evaluative information is already partially activated, resulting in faster response times to the target task. Fazio considered the example of valence in particular. Variants of Fazio et al.'s explanation have been employed by several authors such as Bargh et al. (1988) and Hermans et al. These authors suggest that primes are evaluated in broad terms such as positive and negative, and each evaluation has an individual node in the memory network. Furthermore, all positively valenced objects are interconnected, as are all negatively valenced objects via positively (and respectively negatively) valenced negative nodes. Therefore, activation in one of these nodes makes other concepts linked to this node more accessible. A particular feature of this activation is that it takes place automatically (Hermans et al., 2003; Steinbeis & Koelsch, 2011). In music priming specifically, Steinbeis and Koelsch argue in favour of spreading activation accounts on the grounds that the very short time frame during which affective priming effects are evident suggests that it must be an automatic process.

A second, later, formulation of spreading activation is semantic evaluative activation. Blaison et al. (2012) argue that affective primes suggest evaluative information such as 'good' or 'bad', in common with the affective spreading activation above. The key difference with the explanation above is that the evaluation takes place at a semantic level rather than an affective level. Associative networks partially activate the information associated with these labels, so that the response to a congruent target is facilitated. They argue that affect misattribution is inconsistent with the short SOA associated with affective priming effects. Furthermore, in considering the related affective misattribution theory, Blaison et al.argue that it is semantic misattribution rather than affective misattribution that is the mechanism in play. An additional layer of support for semantic effects is provided by ERP studies, for instance Zhang et al. (2006), where the presence of N400 negativity suggests semantic activation. However, Fazio (2001) argues in favour of a version of semantic priming that activates more specific semantic networks than simply valence; consequently this version of the semantic model would be difficult to apply to music priming. Specifically, this would require some activation of knowledge by a musical stimulus rather than purely affect induction. Furthermore, the results of behavioural tasks that operate on a semantic rather than an affective level, (e.g., Goerlich et al., 2012; Lemmens et al., 2007) suggest that purely semantic explanations of priming effects do not account for affective priming. Moreover, it is intuitively difficult to reconcile a semantic explanation of affective priming with the fact that there is a consistently different pattern of results in the lexical decision task studies from affective categorisation studies. Boakes (2010) attributes the popularity of semantic activation to the superficial similarity in the semantic and affective priming paradigms and their historical connotation.

3.2.2 Response priming

Klinger et al. (2000) and Scherer and Lambert (2009) argue in favour of an action model of priming. In most affective priming tasks, distinct motor responses are associated with positive or negative affect - e.g. a button press or 'a' key for negative, 'l' key for positive. Consequently, exposure to an affective auditory stimulus could pre-activate the corresponding embodied response and so a target with the same affect would result in a faster response than a target with a different affect. To some extent, this explanation is supported by a range of studies that manipulate the frequency of congruent vs incongruent pairs. There is a strong body of evidence (see for instance Freytag et al., 2011) that the frequency of congruent prime-target pairs has a significant impact on response time. A higher instance of congruent prime-target pairs is associated with faster response. This does seem to support the response priming argument reviewed by Winkielman et al. (2007) in that a greater frequency of congruent pairs could feasibly strengthen the priming relationship. However, the frequency studies could also be seen to support the activation arguments in so far as a higher frequency of activations could lead to the information being more accessible.

Next, we consider a competing account of affective priming that is based on Stroop-like interference.

3.2.3 Interference models

A key question in priming paradigms is whether the priming effect (usually the difference in response times between congruent conditions and incongruent conditions) is a consequence of facilitation in congruent conditions, or of interference in incongruent conditions. Dignath et al. (2017) consider the role of valence and arousal in relationship to conflict monitoring mechanisms in the Simon effect. They argue that intrinsically conflicting activation is classed as negative. There is a consequent aversive motivational tendency. This is a critical difference from the interference models proposed by Wentura (2000). Again, the authors emphasise sequential effects. However, despite the theoretical argument they presented, they actually found no significant effect of affect on conflict resolution. Moreover, they found strong Bayesian argument against post-conflict cognitive control. Moreover, although several ERP studies found that incongruent conditions are associated with N400 activity, whereas the conflict monitoring effect is more closely associated with N200 activity (Steinbeis & Koelsch, 2011). Zhang et al. (2006) found that both N200 and N400 activity was elevated under incongruent conditions and suggested that N400 may therefore be associated with conflict monitoring alongside its semantic function. Furthermore, in a purely lexical priming task, Bartholow et al. (2009) found significant N200 activity, lending weight to the conflict resolution concept, but with the caveat that this is applied to a semantic rather than an affective task. Given the relative strength of the ERP evidence in favour of the conflict monitoring, it seems likely that some sort of conflict - whether at the processing or response stage - is, at least in part, responsible for affective priming effects. However, rather than being a reformulation of the action account above, this account places the emphasis on response competition rather than the activation of an embodied response. Goerlich et al. (2012) argue against this on the grounds that on a musically-primed non-affective categorisation task, there was no congruency effect in the negative-negative condition; spreading activation therefore was, in their opinion, a more likely candidate to explain affective priming effects. Zhang et al. assert that their behavioural findings support the idea of Stroop-like interference being responsible for inhibition under incongruent conditions. Indeed, some of the theoretical explanations for facilitation, such as automatic vigilance, are difficult to reconcile in both the positive and negative congruent conditions. An explanation based solely on interference rather than facilitation as a basis of priming effects automatically resolves some of the theoretical inconsistencies present in the other accounts. Moreover, Voss et al. (2013) found that in affective priming, slower reaction times in incongruent conditions are the result of an increased non-decision time compared to congruent conditions, arguing that this favoured an interference account of priming. However, it should be noted that both interference and facilitation effects could be present in priming, i.e., a congruent activation is faster than an unprimed activation, which is in turn faster than an incongruent activation.

3.3 Priming in Music Research

Priming studies have been used in music cognition research since Bharucha and Stoeckig conducted research into harmonic expectation in the 1980s (Bharucha & Stoeckig, 1986, 1987). Despite the paradigm's popularity in other domains, it has been used relatively infrequently in music. Nevertheless, there is a body of research that employs affective priming and related paradigms in music. Here, the literature review will outline the general trends in priming research in music and other auditory domains before discussing more specifically those studies which are particularly important to the present thesis.

In parallel with Fazios introduction of the affective priming paradigm, Bharucha and Stoeckig developed an analogous priming method to investigate harmonic expectation. Although not strictly afffective priming, the studies conducted by Bharucha and collegues form an important part of the evolution of priming studies in music and so are included here. Starting with Bharucha and Stoeckig (1986), they used a unimodal (chord - chord) priming paradigm. Participants classified target chords as being either major or minor; the target chords were immediately preceded by prime chords that were either harmonically related or unrelated to the target. They found congruency effects for major targets, i.e., target chords were classified faster and more accurately when preceded by a related compared to an unrelated prime. However, the effect was absent (or possibly even reversed) when the target chords were minor. In a follow-up experiment, participants determined whether a target chord was in tune or out of tune. Again, the prime chords were either harmonically related or unrelated to the targets. Congruency effects were evident for both major and minor target chords, with decisions being made more quickly and more accurately when primes were related than when they were unrelated. The authors developed a spreading activation account to explain the presence of congruency effects.

Subsequently, the same authors (Bharucha and Stoeckig, 1987) tested their spreading activation account against one competing psychoacoustic explanation based on overlapping frequency spectra. They found that shared spectral components did not influence priming. Again, they concluded that spreading activation provided the best explanation of harmonic priming. Later, Tekman and Bharucha

(1998) carried out a similar study. Critically, they introduced stimulus onset asynchrony (SOA) as an additional variable. They found that at short SOA (i.e., 50 ms), priming effects were governed by sensory i.e., psychoacoustic components, whereas at longer SOA (500 ms) the effects were governed by cognitive components, namely harmonic convention. It should, however, be noted that there are other psychoacoustic features (roughness, sharpness, spectral flux) that were not tested. The same authors had previously carried out a more comprehensive time-course analysis (Tekman and Bharucha, 1992), which did not consider alternative mechanisms, and established the presence of harmonic priming at SOA = 50 ms to SOA = 2500 ms. Indeed, they found that there was no variation in the size of the effect at different SOA.

Justus and Bharucha (2001) probed the competing roles of long-term (schematic) knowledge versus short-term (veridical or local) knowledge. They found that congruency effects were predicted by the schematic knowledge, rather than by the short-term learning of veridical information. Again, this provides evidence in favour of the cognitive rather than the sensory mechanism. However, it should be noted that this study used relatively long SOA (1000 ms), and was outside the window for the sensory mechanism to have primacy.

Bigand et al. (2003) probed this further, with a more comprehensive time-course analysis comparing cognitive and sensory components. They found that at 300 ms and 150 ms, cognitive components drove priming. However, at SOA = 75 ms, cognitive components were implicated only if participants had been exposed to the other conditions first: if they had not been exposed to the 300 ms and 150 ms conditions, participants responses were dictated by sensory components.

Johanis and Schmuckler (2021) adapted the tuning paradigm employed by Bharacha and Stoeckig (1986), extending it into cross-modal phenomena. Using a participant pool of trained musicians, they found that visually presenting the musical notation of a triad, plus sounding the root of the triad was sufficient to replicate Bharucha and Stoeckig's (1986) finding that tuning judgements were slower and less accurate for harmonically unrelated compared to related chords, whereas presenting the notation without sounding the root did not result in replicating the findings.

The discussion will now turn to affective priming studies of the sort carried out in the present thesis. Although the affective priming paradigm is in general terms well-researched, the number of music studies that use this method in the strict sense is relatively small compared to other domains such as linguistics, memory or social cognition. The first study, to the present author's knowledge, to use the 'standard' affective priming paradigm in music was Sollberger et al. (2003). This study considers the role of consonance and dissonance and modality (major vs minor) in affective priming. This study proved influential in providing a prototype for later studies. As well as focusing on single chord primes with a focus on clear polarities such as consonance-dissonance or major-minor, other methodological details, particularly the 200 ms SOA, were used by authors such as Costa (2013) and Huziwara et al. (2023). Likewise, in a pair of related publications, Steinbeis and Koelsch (2008, 2011) considered the role of consonance and dissonance and modality in affective priming. Steinbeis and Koelsch built incrementally on the work of Sollberger et al. in probing the role of musical training in priming, as well as more precisely defining consonance and dissonance in terms of roughness. Moreover, both Steinbeis and Koelsch articles considered ERPs in addition to reaction time and accuracy measures. Costa (2013) replicated the findings of Sollberger et al. and Steinbeis and Koelsch when consonant and dissonant chords were used to prime responses to word targets. Costa et al. added an additional manipulation by varying the pitch of the prime chords. Costa also probed the target space by replacing the target words with target images. Curiously, they found that the consonance-dissonance manipulation did not lead to congruency effects in reaction times or accuracy rates when evaluating target images (in contrast to evaluating target words, where congruency effects were present). However, congruency effects were present when pitch (high vs low) was the manipulation of interest – that is, positive images were evaluated more quickly when preceded by high pitched chords compared to low-pitched chords, whereas the converse was true of negative target images: they were evaluated more quickly when preceded by a low-pitched chord. Costa explains the failure of consonance and dissonance to prime responses to positive and negative images by suggesting the higher arousal level induced by the target images compared to the primes overrode any transfer of affect from the primes. However, the arousal explanation does not provide a clear account of why pitch should function effectively as a prime. Moreover, Carroll and Young (2005) found that environmental sounds were capable of priming responses to facial image targets.

An alternative explanation for congruency effects that was not posited in Costa (2013) could be found in the acoustic properties of the prime stimuli. For instance, Steinbeis and Koelsch use roughness as the measure of consonance-dissonance. It is plausible that it is not pitch per se that is responsible for the presence of congruency effects, but possibly that owing to the widened (at least of the CB is determined in terms of the musical interval) critical band at lower pitches that higher roughness levels for chords that are otherwise the same except for pitch could be responsible for the congruency effects. Lahdelma, Armitage, and Eerola (2022) also consider consonance and dissonance as the key manipulation in priming evaluative responses to valenced words. In addition, they also consider possible roles for numerosity (i.e., how many notes there are in a particular chord) and timbre (at least in the limited case of harmonium compared to piano.) Like the authors cited above, they found the predicted congruency effects. However, this was qualified by an interaction with numerosity – they found that congruency effects were not detected when two note chords (i.e., intervals) were used as primes; on the other hand, strong congruency effects were detected when four-note chords were used as primes. Timbre manipulation was not found to influence congruency effects. The authors tentatively speculated that numerosity was in fact a proxy for roughness. This line of reasoning is taken up in Chapter 9. The final study in this line is Huziwara et al. (2023), which replicates many aspects of Costa (2013) using emotional faces as targets rather than non-facial valenced images. Again, in common with Costa et al., the manipulations under consideration were consonance-dissonance and high-low pitch. Huziwara et al. found congruency effects with the exception of high-pitch happy face condition. The authors contend that the salience of the happy face reduced the impact of the prime. There is a small number of additional studies that used single chords as primes. Zhao et al., 2019 used major-minor and consonant-dissonant chords as primes with happy-sad facial expressions. Critically, this study considered group differences between participants with amusia and typically-developed controls. Both manipulations, major-minor and consonance-dissonance, elicited congruency effects in the control group but not the amusia group. However, it should be noted that the consonance-dissonance manipulation was defined in terms of harmonicity whilst roughness was held constant. Marin et al. (2015) found that, in a rating task, amusics were sensitive to roughness but not harmonicity. It is plausible that the sensitivity reported with self-report ratings could be mirrored in an affective priming task had the manipulation been roughness rather than harmonicity. Zhao et al. (2019) found priming effects with consonant-dissonant chords and happy-sad facial targets. There are two final studies that uses single-chord stimuli as primes, Ragozzine (2011) and Bakker and Martin (2015), but which differ slightly from the previously mentioned studies in some methodological respects. Ragozzine's first experiment found no priming effects at a long stimulus onset asynchrony with major-minor chords as primes, but the second experiment with a shorter (250 ms) SOA did find congruency effects. Bakker and Martin did not find congruency effects using single-chord primes. However, they employed an SOA of 0 ms – in the strictest sense, simultaneous presentation could exclude this study from being considered a 'true' priming study.

In addition to the studies that use single chords as primes, three studies use ecologically valid music clips[1] as primes. Lense et al. (2013) considered group dif-

[1]From this point onward, the terms 'music prime' and 'chord prime' will be used to distinguish between studies that use individual chords as primes and those that use evolving music clips as primes.

ferences between particidonepants with Williams Syndrome (a neurodevelopmental disorder that is characterised by heightened emotion recognition as well as an interest in music) and typically-developed controls. They found that congruency effects were not present in the control group but were present in the Williams syndrome group. However, the generalisability of Lense at al.'s result is potentially limited by the use of an SOA of 750 ms, which is considerably longer than other music and priming studies. Goerlich et al. (2012) found a marginal congruency effect using music primes. Moreover, Tenderini et al. (2022) also found congruency effects with both native speaker and second language participants in an affective priming task using music primes.

A small number of studies, (Goerlich et al., 2011; Lense et al., 2013; Zhou et al., 2019) consider priming as a vehicle to consider individual differences (Williams syndrome, amusia, and alexithymia respectively.) As outlined above, Lense et al. found congruency effects in participants with Williams Syndrome, but not in typically-developed controls, albeit with a long SOA, and Zhou et al. found that priming was present in typically developed controls, but not in amusics. Goerlich et al. found that there was a negative correlation between the degree of alexithymia exhibited by a participant and the size of their N400 response to incongruent conditions.

3.4 Meta-analysis of relevant studies

Although music priming studies of the type discussed here have been carried out for some twenty years, there has up until now been no attempt to review or summarise all of the published articles in the field. In particular, although Herring et al. (2013)

have carried out a comprehensive meta-analytic review of the field more broadly, and there are a number of narrative reviews of affective priming, music priming studies fall outside the scope of previous reviews. Consequently, the present meta-analysis attempts primarily to summarise the state of the art in music priming and to fill the gap in providing a comprehensive summary of music priming research to date. In addition, it seeks to consider moderator variables that may influence the size of the priming effect, such as prime type, target type, stimulus onset asynchrony (SOA) and total number of trials. A meta-analysis provides a framework to explore the literature systematically, and derive an estimate for the size of the priming effect across a broader sample of participants.

Table 3.1 summarises music priming studies (alongside some studies which consider non-musical auditory stimuli).

3.4.1 Search

Appropriate studies were identified by searching Scopus, PsychINFO and Web of Science using the following keywords: *affective priming music, consonance dissonance priming, evaluative priming music*

58 unique articles were identified by this method. These articles were then subjected to the inclusion-exclusion criteria outlined below.

3.4.2 Inclusion-exclusion criteria

Studies that used music primes and non-music (e.g., picture or word) targets were included. Studies that used music as targets are excluded from the analysis. The classification tasks had to be a keyboard response in which targets were identified as positive or negative; the dependent variable was restricted to reaction time or accuracy rate. ERP studies which reported behavioural data were included, though the ERP data was not considered as part of the meta-analysis. Finally, as affective priming is considered to be an automatic effect, studies with a stimulus onset asynchrony of more than 1000 ms were excluded as any effects found by such studies are likely to be the consequence of conscious processes. Two studies were excluded because they reported the same data as another study by the same authors. Published articles that form part of the present thesis were excluded a priori and so have not been included in the meta-analysis. However, a related study, which does not form part of the thesis (Lahdelma, Armitage, & Eerola, 2022) is included in the analysis. All studies were published in English in peer-reviewed journals. Unreviewed and unpublished material was not included.

3.4.3 Studies

58 unique studies were identified via the initial search. Applying the inclusion-exclusion criteria reduced the numbers as follows:

- not priming study (10 removed)
- non-music studies (13 removed)
- SOA 0 ms or >1000ms (10 removed)
- music targets (5 removed)
- present thesis (4 removed)

- not peer reviewed (2 removed)
- data reused in later publications (2 removed)

The following studies were identified as being suitable for inclusion in the meta-analysis: Costa, 2013; Goerlich et al., 2011; Huziwara et al., 2023; Lahdelma, Armitage, and Eerola, 2022; Lense et al., 2013; Ragozzine, 2011; Sollberger et al., 2003; Steinbeis and Koelsch, 2008, 2011; Tenderini et al., 2022; Zhao et al., 2019; Zhou et al., 2019. Where articles consisted of multiple experiments, the experiments are reported separately in the present meta-analysis. Where studies used groups where the research question predicted group differences owing to a neurodevelopmental or congenital disorder (Lense et al., 2013; Zhou et al., 2019), only results for the typically-developed control group are included. Overall, the meta analysis considered responses from 598 participants.

3.4.4 Calculation of effect sizes

The critical relationship is the relationship between prime valence and target valence. In three studies this was defined as congruence, in nine studies this was defined by the interaction of target valence and prime valence. Following the meta-analytic procedure used by Herring et al. (2013), a priori, it was determined that repeated measures designs would be reported using Cohen's *d*. Cohen's *d* provides a standardised measure of difference in mean reaction time between congruent and incongruent conditions. Where reported by the authors, effect sizes are given here as reported. When Cohen's *d* is not reported by the authors, or where insufficient information has been reported to calculate *d* directly, it is estimated here using the formulae (Morris & DeShon, 2002):

$$d = \sqrt{\frac{F}{n}} \text{ and } \frac{t}{\sqrt{n}}$$

.

In repeated-measures designs, such as those investigated here, Cohen's d can be estimated more accurately using the correlation between reaction times in incongruent and congruent conditions. However, since this is not reported in any of the studies included in the present meta-analysis, the above formulae were used.

In addition, it was necessary to calculate the variances in the effect sizes. Again, to be calculated exactly, the correlation between in congruent and incongruent conditions is necessary. Therefore, the following approximation was used to estimate the standard error in effect sizes:

$$SE^2 = (\frac{1}{n})(\frac{n-1}{n-3})(1+nd^2) - \frac{d^2}{[c(df]^2} \text{ where } c(df) = 1 - \frac{3}{4df-1}$$

Between-study heterogeneity indexes the degree of variation in true effect sizes between the studies reported in a meta-analysis. In order to determine the degree of heterogeneity between the studies, a Cochran's Q test was carried out on the effect sizes. This tests whether or not the mean effect size represents a common population mean. This proved significant, $Q(17) = 999.95$ $p < .0001$. As an additional measure of heterogeneity, the I^2 index (Higgins & Thompson, 2002) was also calculated, $I^2 = 98.3\%,\ 95\%\,CI = [97.9, 98.6]$. This estimates the percentage of variability in the results between the included studies that is a consequence of genuine differences in the results rather than random variation. However, I^2 has been subject to criticism, particularly when the number of included studies is small, as is the case here. The final measure of between-study heterogeneity was estimated as $\hat{\tau}^2 = .0878$ $95\%\,CI = [0.0476, 0.2085]$. Overall, the seemingly high degree of heterogeneity suggests that a fixed-effects model is not appropriate (Hedges & Vevea, 1998) and so a mixed effects model has been chosen to perform the meta-analysis (Hedges, 1992). The mixed effects model has the benefit of considering both fixed effects (i.e., that there is one population effect size) with random effects (which attribute differences in effect size to both sampling error and variation between studies).

Publication bias – i.e., publishing only studies with clear effects – has been well-reported in psychological disciplines, and is known to influence the results of meta-analyses (Thornton & Lee, 2000). In order to measure bias, the studies were plotted on a funnel plot. Funnel plots feature a measure of effect size on the x-axis with a measure of the studies' precision, e.g. sample size or standard error on the y-axis. Studies that are high in precision are close to the average x-value; studies that are low in precision are plotted further from the average. A symmetric plot indicates low publication bias whereas an asymmetric plot indicates possible publication bias. A number of possible moderating factors that may have influenced the effect sizes are introduced below.

Prime type

The majority of music priming studies (Costa, 2013; Huziwara et al., 2023; Lahdelma, Armitage, & Eerola, 2022; Ragozzine, 2011; Sollberger et al., 2003; Steinbeis & Koelsch, 2008, 2011; Zhao et al., 2019; Zhou et al., 2019) consider single chord stimuli, whereas Goerlich et al. (2012), Lense et al. (2013), and Tenderini et al. (2022)

consider short musical extracts. The manipulations for single chords have included combinations of mode (major vs minor), consonance (consonant vs dissonant), pitch (high vs low), timbre (piano vs harmonium). However, the operationalisation of consonance-dissonance has been inconsistent with roughness (Steinbeis & Koelsch, 2008, 2011), harmonicity (Zhao et al., 2019) and perceived pleasantness (Sollberger et al., 2003) all used as indices of consonance-dissonance. Musical stimuli have typically been taken from previous literature and pre-rated for valence. Lahdelma, Armitage, and Eerola consider numerosity alongside consonance-dissonance and timbre as an additional manipulation.

Target type

Studies with music primes have typically used visually presented stimuli as targets, specifically valenced words (Costa, 2013; Goerlich et al., 2012; Lahdelma, Armitage, & Eerola, 2022; Steinbeis & Koelsch, 2008, 2011; Tenderini et al., 2022), happy-sad faces (Huziwara et al., 2023; Lense et al., 2013; Zhao et al., 2019), or non-facial images (Costa, 2013). Typically, target stimuli are either selected from pre-existing databases of affective stimuli (e.g., Costa, 2013; Lahdelma, Armitage, & Eerola, 2022) or pre-rated by a separate panel for validation as use as stimuli (e.g., Goerlich et al., 2011). In general, there is consensus that word targets elicit priming effects. However, the situation is less clear with images. Costa and Huziwara found no evidence of priming when images were used as primes, whereas Zhao and Zhaou did report priming effects with image (facial) targets. Lense et al. did not find evidence of priming with faces used as targets with a typically-developed control group. Curiously, the authors did find evidence of priming with a group of participants with

Williams syndrome. However, it is possible that the long gap (750 ms) between the onset of the prime and target stimuli may have had an impact on the results (see below). Costa argues that the absence of priming with image targets is due to the increased arousal conveyed by the images compared to word targets. However, pitch was found to prime image targets (Costa, 2013; Huziwara et al., 2023).

Stimulus Onset Asynchrony

Stimulus onset asynchrony (SOA) is the gap between the onset of the first stimulus (i.e., the prime) and the second stimulus (i.e., the target). In the priming literature more broadly, there is a consensus that priming effects are sensitive to SOA. In particular, priming effects are maximised when SOA is in the range 200 - 300 ms and decays thereafter (Hermans et al., 2001). However, it is important to note that most priming studies have used visual presentation of words or images as both primes and targets. To date, there has been no systematic probing of SOA in the auditory domain. However, most authors have used an SOA of 200 ms in conjunction with a relatively simple prime type (single chords; Costa, 2013; Huziwara et al., 2023; Sollberger et al., 2003; Steinbeis and Koelsch, 2008, 2011). Strikingly, Ragozzine (2011) found that, with major/minor chords as primes, there were no priming effects with an SOA of 1000 ms but that when the SOA was reduced to 250 ms there was a significant priming effect. For ecologically valid music samples, Goerlich et al. (2011) and Tenderini et al. (2022) found priming effects with an SOA of 200 ms, whereas Lense et al. (2013) did not find priming effects at an SOA of 750 ms. Using environmental sounds as auditory primes, Scherer and Larsen (2011) found priming effects at an SOA of 450 ms, arguing that the slightly longer than customary SOA was important as auditory signals evolve over time in contrast to static visual primes. In contrast, Bakker and Martin (2015) found that at SOA = 0 ms (i.e., with simultaneous presentation of prime and target – and hence not strictly speaking *priming*) there was no evidence of affective priming. Conversely, Degner et al. (2012) considers the size of congruency effects at SOAs of 0 ms, 250 ms and 500 ms. They found that priming effects were greater at SOA of 0 ms compared to 500 ms, suggesting that priming was present only if there was some overlap in the presentation of the stimuli. Overall, there seems to be consensus that stronger congruency effects are detected at shorter SOA.

Number of trials

There is a large degree of variation in the number of trials in the literature, ranging from twenty-eight trials per participant (Ragozzine, 2011) to 256 trials per participant (Lahdelma, Armitage, & Eerola, 2022). Although there are differing conventions on the optimal strategy to optimise power for repeated measures designs with multiple readings per participant per condition (see e.g., Brysbaert & Stevens, 2018, for discussion), Brysbaert and Stevens suggest a minimum total of 1600 readings per condition to achieve a power of 0.8 with $\alpha = .05$. Number of trials was found to be a significant predictor of effect sizes in a previous meta-analysis of non-music affective priming studies (Herring et al., 2013). In keeping with the approach taken by Brysbaert and Stevens, in this meta-analysis, the total number of trials is defined as the number of trials per condition summed over all participants. It is calculated here as $\textit{number of trials} \times \textit{number of participants}$.

Musical training

A number of studies include musical training as a potential group difference. Steinbeis and Koelsch (2008, 2011) did not find any significant group differences when consonant and dissonant chords were used as primes. Although there was no significant group difference when major and minor chords were used as primes, Steinbeis and Koelsch did report a non-significant trend towards a greater priming effect being present in musicians when these stimuli were used. Moreover, Lahdelma, Armitage, and Eerola (2022) did not find an effect of musical training. It seems plausible that musical training may influence susceptibility to priming effects when the affective nature of the prime is defined by culture (e.g. the major-minor distinction) whereas prime features that are dictated by psychoacoustic factors (such as roughness) are not susceptible to the influence of musical training. Some studies use participants without musical training (e.g., Costa, 2013), others probe group differences specifically (Steinbeis & Koelsch, 2011) and others report musical training incidentally (Lahdelma, Armitage, & Eerola, 2022; Ragozzine, 2011). However, given the unanimity of findings against group differences between musicians and non-musicians, it has not been included as a possible mediating variable in the present analysis.

Table 1 contains a summary of affective priming studies that use music primes.

Table 3.1: Summary of music priming studies

Author	Year	Primes	Targets	SOA	Findings	Cohen's d	n
Sollberger et al. Exp 1	2003	Consonant vs dissonant chords (3/4 notes)	Words	200 ms		0.71	42
Sollberger et al.	2003 Exp 2	Consonant vs dissonant chords (4 notes)	Words	200 ms		0.40	76
Steinbeis and Koelsch Exp 1†	2011	Consonant vs dissonant chords	Words	200 ms		0.82	40
Steinbeis and Koelsch Exp 2	2011	Major vs minor chords	Words	200 ms		0.62	40
Ragozzine Exp 1	2011	Chords	Words	1000 ms		0.23	12
Ragozzine Exp 2	2011	Chords	Words	250 ms		0.88	10
Goerlich et al.	2011	Music	Words	200 ms	marginal effect of congruence	0.32	32
Costa Exp 1	2013	Consonant vs dissonant chords × high vs low chords	Words	200 ms		0.40	70

Costa Exp 2	2013	Consonant vs dissonant chords × high vs low chords	IAPS images	200 ms	C/D ns/High-Low sig	0.01	41
Costa Exp 3	2013	Major vs minor chords *times* High vs low chords	IAPS images	200 ms	Major/minor ns/High-Low sig	0.00	41
Lense et al. Controls only	2013	Valenced instrumental music	Faces	750 ms	congruence non-significant	0.00	26
Zhou et al. Exp 1 controls only	2019	Major vs minor	Happy vs sad faces	200 ms		0.42	16
Zhou et al. Exp 2 controls only	2019	Consonant vs disssonant chords	Happy vs sad faces	200 ms		1.00	16
Zhao et al.	2019	Consonant vs dissonant chords	Happy vs sad faces	200 ms		0.57	15
Lahdelma, Armitage, and Eerola	2022	Consonant vs dissonant two and four-note chords	Words	200 ms		0.57	40
Tenderini et al. Exp 1	2022	Valenced music	Words	200 ms		0.61	31

Tenderini et al. Exp 3	2022	Valenced music	Words	200 ms		0.89	23
Huziwara et al.	2023	Consonant/dissonant chords *times* pitch (high/low)	Happy/sad faces	c/d*face and pitch*face both sig	200 ms	0.71	40

† results from Steinbeis & Koelsch (2008) are subsumed into Steinbeis & Koelsch (2011)

Data analysis

The meta analysis was carried out in accordance to the protocol laid out in Forero et al. (2019), Harrer et al. (2021), and Page et al. (2021). Data analysis and visualisation for the meta-analysis was carried out in R (R Core Team, 2023) using the packages *mata* and *dmetar* (Balduzzi et al., 2019; Harrer et al., 2019). Data, analysis scripts and further details of the inclusion-exclusion procedure are available at https://osf.io/t4rf3/?view_only=2e744ef545df4bbfad8a0df4503c86f2.

Eighteen effect sizes were considered from experiments in twelve published articles. All experiments used evaluative decision tasks where the response variable was reaction time. The mean SOA was 282 ms (SE = 55.23). Nine articles (75%) used chords as prime stimuli whereas three used musical extracts (25%). For the studies that used chords as stimuli, fourteen effect sizes were considered, where the manipulations were major-minor (six effects, 32%), consonance-dissonance (nine effects, 47%) and pitch (high-low; four effects, 21%). Eleven effects (48%) related to word targets, twelve (52%)related to image targets (either general valenced images, or happy-sad faces). The mean number of trials was 95.1 (SE = 10.9) and the mean number of participants was 33.8 (SE = 4.7). The weighted mean effect size was calculated as $d = 0.50,\ SE = 0.03\ 95\%CI = [0.36, 0.64]\ z = 5.06\ p < .001$. Next, a series of moderator analyses is carried out for the purpose of accounting for the heterogeneity. Figure 3.2 below shows the weighted effect sizes, d_i and the 95% confidence intervals.

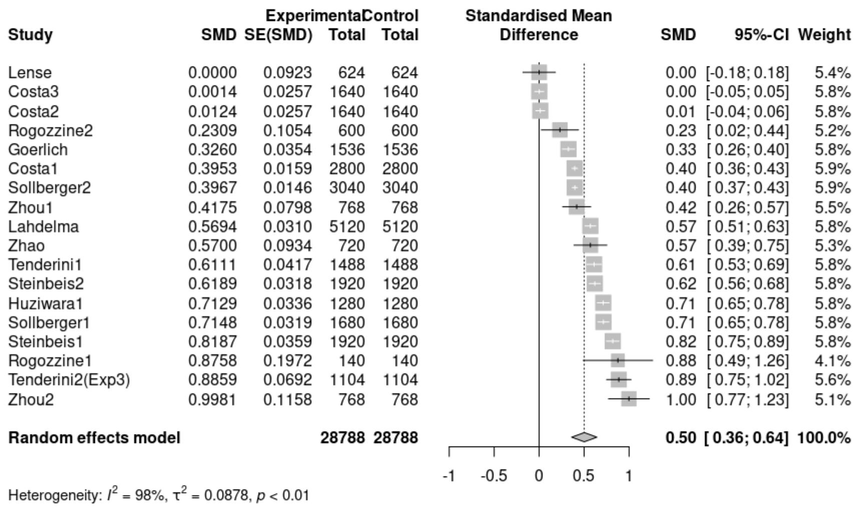


Figure 3.2: Forest plot of the 18 effects from 12 articles. The weighted mean effect size is shown by the diamond at the bottom.

3.4.5 Moderator analyses

The relationship between prime-type, manipulation and target type was investigated using a sequence of meta-regressions using the R package Balduzzi et al., 2019. The initial model tested was $d \sim PrimeType \times TargetType \times Number\ of\ Trials$. SOA was tested separately as only two studies had high SOAs and so it was not possible to isolate its effect from other variables. SOA was dichotomised and the effect sizes for low SOA and high SOA (defined as less than 500 ms and greater than 500 ms respectively) were compared.

The results of the regression analysis are summarised in Table 3.2.

	β	$SE\ \beta$	Z-value	p-value	95% CI β
Target	-0.6257	0.5230	-1.1963	0.2316	-1.6508 0.3994
Prime (Major-Minor)	-0.5882	0.7061	-0.8329	0.4049	-1.9722 0.7958
Prime (Music)	-0.2843	0.6181	-0.4600	0.6455	-1.4957 0.9271
Number of readings	-0.0004	0.0002	-2.0322	0.0421	-0.0007 -0.0000 *
Target x Prime (Major-Minor)	0.3968	0.8172	0.4856	0.6273	-1.2048 1.9984
Target x Prime (Music)	1.6261	0.7633	2.1304	0.0331	0.1301 3.1222 *
Target x Number of readings	0.0003	0.0002	1.8048	0.0711	-0.0000 0.0007
Prime (Major-Minor) x Number of readings	0.0001	0.0003	0.4404	0.6596	-0.0004 0.0007
Prime (Music) x Number of readings	-0.0005	0.0004	-1.2599	0.2077	-0.0013 0.0003
Target x Prime x Number of readings	-0.0001	0.0003	-0.3183	0.7502	-0.0007 0.0005

Table 3.2: Summary of meta-regression results. The reference level is consonant vs dissonant primes with image targets. Significance levels are indicated as follows: *** p<0.001; ** p< 0.01; *p< 0.05

Considering the results in Table 3.2, we see that the number of readings, a proxy for power (Brysbaert & Stevens, 2018), is a significant predictor of effect size. The negative value of β suggests that as the number of trials increases, the effect size decreases. There is also a significant interaction of prime type and target type: with word targets, prime type does not influence effect size, whereas with image targets, there are significantly lower effect sizes with music primes as compared to consonant and dissonant chords. This is represented graphically in Figure 3.3

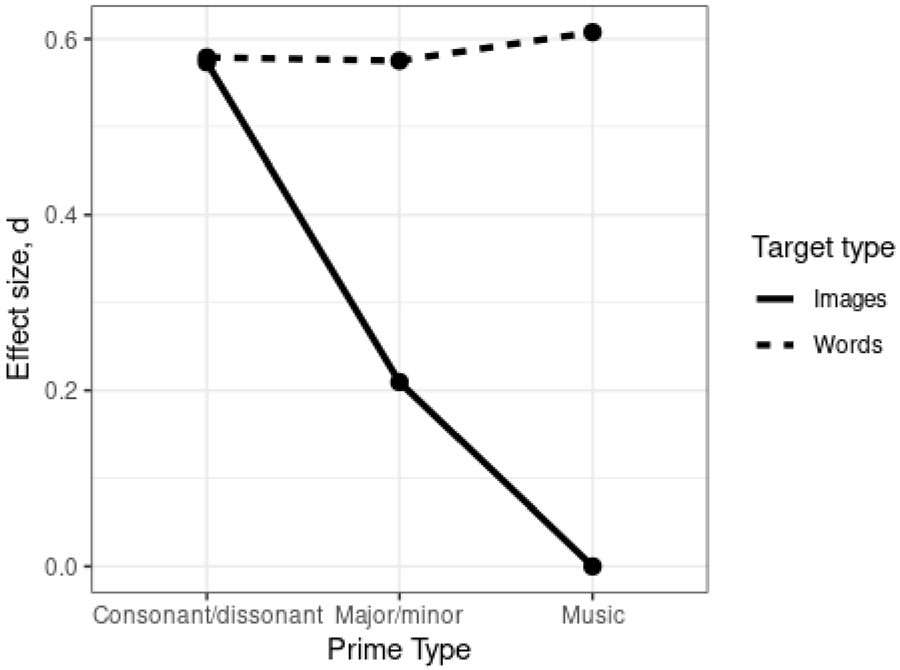


Figure 3.3: Effect of prime type and target type on effect size, d

SOA was coded such that SOA<500 ms was categorised as low, whereas SOA>500 ms was categorised as high. Effect sizes associated with a low SOA value (mean $d =$ 0.558) were significantly greater than those associated with a high SOA value (mean $d = 0.115$), $\beta = 0.4321,\ 95\%\ CI = [0.0054, 0.8588],\ z = 1.9846\ p = 0.0472$.

Publication bias

One of the key challenges facing psychological science in recent years has been publication bias (see, e.g. Francis, 2012). That is, a tendency for statistically significant results to be published but for non-significant results to remain unpublished. This in turn creates a challenge for any attempt to systematically analyse data from across a domain, as the published literature may not be representative of the findings of the body of (potentially unpublished) research as a whole. Thus, options for researchers carrying out meta-analyses are to 1) contact authors active in the field and request copies of unpublished data, or 2) employ an analytic strategy (Egger et al., 1997) to identify the extent of any publication bias. Here, the latter course of action has been chosen. Figure3.4 – a funnel plot – shows the relationship between the standardised mean difference (i.e., Cohen's *d*) and the total number of trials, a proxy for power (Brysbaert & Stevens, 2018). The dotted vertical line indicates the unweighted mean effect size, and the solid vertical line indicates the weighted mean effect size. Typically, in the absence of bias, studies would fall randomly either side of the weighted mean, with a wider range of effect sizes for lower powered studies, and with higher powered studies nearer to the weighted mean, broadly following the contours of the funnel. That is to say symmetric funnel plots indicate an absence of bias, whereas asymmetric plots that do not conform to the funnel are considered to be indicative of *possible* bias.

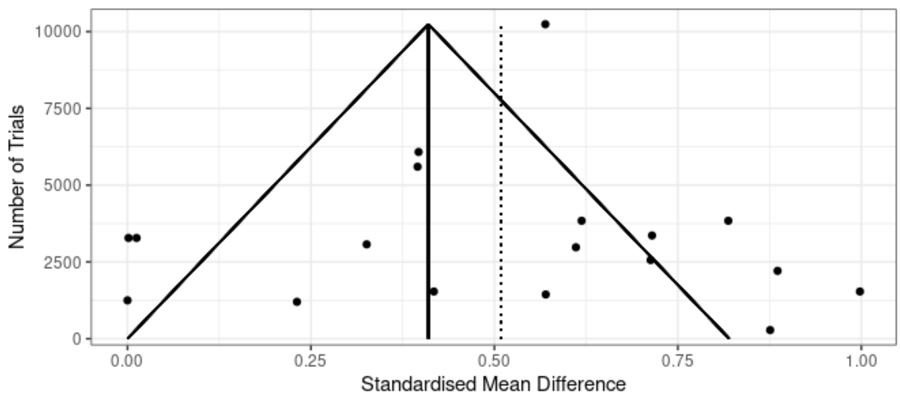


Figure 3.4: Funnel plot of standard errors plotted against effect sizes

Visual inspection of Figure3.4 suggests that there is a possibility of publication bias, as indicated by the asymmetry in the plot. In particular, there is a cluster of five relatively low-powered studies in the bottom right of the funnel plot.

As graphical tests are not considered conclusive evidence of bias (Hunter et al., 2014), to investigate the possibility of publication bias further, it was necessary to employ an analytic test. Bias is frequently tested using Egger's regression test (Egger et al., 1997). In this instance, Egger's test proved non-significant, intercept $= 3.611$, $95\%\ CI = [-2.92, 10.14]$, $t = 1.083$, $p = .29$. This suggests that there is no evidence of publication bias in the studies included in the meta-analysis, even though there is a tendency towards reporting stronger findings.

3.4.6 Discussion

Chapter 3 has summarised the music priming literature from the last 20 years. There is a strong consensus that congruent conditions facilitate reaction time whereas there is inhibition in incongruent conditions, although there is a degree of variablity in the size of the effect. However, it should be noted that a small number of studies do not accord with the consensus view (Costa, 2013; Lense et al., 2013), although this may be explained by mediating factors which will be discussed further below.

The meta-analysis revealed significant effects of the moderating factors SOA and number of readings. The music priming literature is in a broad consensus with the priming literature in general in finding priming effects up to SOA = 300 ms, with the effect decaying at longer SOA. Curiously, this is at odds with Degner et al.'s (2012) time course analysis, which found optimal affective priming at SOA = 0 ms for auditory - auditory priming. Indeed, this is in keeping with the time course analysis carried out by Hermans et al. (2001). However, there are at least two instances of auditory affective priming being detected at slightly longer SOA: 450 ms in the case of Scherer and Larsen (2011, using environmental affective sounds), and at 750 ms in the case of Lense et al. (2013, using music primes with a group of participants with Williams Syndrome). In general, low SOA is favoured in the priming literature. (It should be noted that SOA - 0 ms is considered by some authors (for instance Herring et al., 2013) to be a phenomenon different from priming as priming is predicated on the prime stimulus being presented *before* the target, and not simultaneously.) There was also an effect of the number of readings – a better measure of power than simple sample size (Brysbaert & Stevens, 2018). The effect size reduced as the number of trials increased, suggesting that studies with a small number of readings inflate the effect size. Indeed, Figure 3.4 reveals a cluster of studies with a small number of readings but large effect sizes. This also suggests that, as the power of a study of this design is dictated by the number of 'readings per condition' (Brysbaert & Stevens, 2018), a number of studies are arguably underpowered, either as a result of a small sample sizes or of a small number of trials. Consequently, it seems plausible to question the extent to which the consistency of the presence of effect sizes is reflective of an underlying consistency or if it is related to the lack of power of some of the studies.

Finally, there was an interaction of prime type and target type that indicated that, whilst prime type did not influence effect size when the targets were words, prime type did influence the effect size when the targets were images. In particular, music primes were associated with lower effect sizes than when consonant/dissonant chords were employed as primes. Although some authors, notably Costa (2013) and Huziwara et al. (2023), suggest that classifications of target images are not influenced by consonant or dissonant primes, the accumulated evidence seems to indicate that priming effects are present with consonant/dissonant chords independent of target type, with faces, non-facial images and words all being associated with priming effects. However, Costa et al's and Huziwara et al.'s results do suggest that the specifics of target images as opposed to words should be researched further. In particular, consideration should be given to the discrepancy between the pitch and consonance-dissonance manipulations reported by Costa et al. A limitation of this result is that there is a possible confound in the music prime - image target condition in that Lense et al. (2013) also employs a long SOA, which is typically associated with smaller effect sizes.

Most music priming studies have used relatively simple primes, e.g. major vs minor or consonant vs dissonant chords, although a small number have used music extracts as primes (Goerlich et al., 2011; Lense et al., 2013; Tenderini et al., 2022). Both prime types, single chords and music clips, have been consistently effective in generating priming effects, Similarly, the specific manipulation – consonant vs dissonant chords, major vs minor chords, happy vs sad music clips – does not seem to determine the presence or otherwise of priming effects, except where related to the interaction of prime type and target type.

A limitation of the present meta-analysis is that it considers congruency as a whole – i.e., positive prime/positive target is considered to be the same condition as a negative prime/negative target. It may be the case that a more granular approach that considers the data as 2 (prime: positive vs negative) x 2 (target: positive vs negative) designs may be more informative in revealing what specific stimulus properties are responsible for priming. In particular, authors outside music cognition have reported a range of results pertaining to negative stimuli, for instance Maier et al. (2003). Additionally, the number of existing studies in music priming is still relatively small. It is also possible that there are other factors that have not been considered as mediators, such as native language, individual differences, headphone use, instructions and awareness of the objective of the task. In addition, it is possible that some aspects of the data analysis strategy may influence the results, e.g., outlier cutoffs or decisions to transform reaction time data.

Overall, a shortcoming of the extant music priming literature is that there is relatively little discussion of the cognitive mechanisms that are engaged in affective priming. Outside of the music priming literature, several authors have discussed the functions of memory, attention and emotion in affective priming (see e.g., Fazio et al., 1986; Hermans et al., 2003; Klauer, 1997; Spruyt et al., 2009), and made attempts to evaluate the relative contributions of spreading activation or Stroop interference accounts of priming. For instance, explanations such as feature-specific attention allocation (Spruyt et al., 2012) have not been considered in music-related studies.

In addition, the fundamental question of how a stimulus in one modality can influence judgements about stimuli in another has received relatively little attention. The differential role of valence and arousal has been considered in a primed rating task (Marin et al., 2012), but not in a reaction time task. A further limitation of the music priming literature is that there has been little consideration given to the extent to which priming with musical primes is similar to or different from priming in other modalities. To date, music priming studies have focused on demonstrating the existence of priming effects and, to a lesser extent, on determining the circumstances under which those effects can be found. For instance, with regard to priming with chords, further research is necessary to establish precisely which stimulus features, e.g., numerosity, roughness, or harmonicity, are implicated in the presence or otherwise of priming effects. Finally, although there have been valuable moves forward in linking certain traits to music priming, (for example Goerlich et al., 2011; Lense et al., 2013, consider alexithymia and Williams Syndrome respectively), the field would benefit from probing a wider range of traits and linking them specifically to the mechanisms involved in priming.

CHAPTER 4

Online Data Collection in Auditory Perception and Cognition Research: Recruitment, Testing, Data Quality and Ethical Considerations

Abstract

Online studies using recruitment services (such as Prolific or Amazon's MTurk) and online testing platforms (such as Gorilla or PsyToolkit) are becoming increasingly common in psychological science. Although auditory disciplines have been slower to adopt these methods, uptake is rapidly increasing in auditory perception and cognition research. Utilising online data collection and recruitment presents several advantages to researchers in terms of the speed of research and the range of target demographics available compared to either traditional lab studies or web-based recruitment via traditional means. Online platforms and recruitment services also present a set of technical and ethical challenges owing to the fact that the people completing experiments are working with their own devices from their homes. This article discusses the potential technical and ethical implications of online studies, including both recruitment services and online testing platforms, with specific reference to auditory perception and cognition research. Rates of remuneration, sampling characteristics, anonymity, quality control, and ethics are all discussed with respect to these approaches. We also provide proposals for how researchers can ensure that online research meets present-day ethical and technical guidelines as well as research transparency standards.

4.1 Introduction

Empirical research in auditory disciplines has historically taken place in highly controlled lab settings. In these settings, researchers have been able to control the exact environment under which participants hear and respond to auditory stimuli, for instance volume, sound quality, distance from speakers or screens, and minimized or completely removed ambient noise and ambient sound distortions and reflections. During the last decade, researchers have begun to collect auditory data online. In part, this change in data collection procedure has been driven by necessity (i.e., the Covid-19 pandemic), but it is also a consequence of increased access to Internet-enabled technologies and the improvements in web-based data collection software. Online data collection technologies can be (and are) used with samples recruited in a traditional way, such as psychology undergraduates meeting course requirements, volunteers from the general public or specifically targeted populations. However, the research community in cognitive sciences, social sciences, economics and AI research is increasingly turning to recruitment services to target large numbers of participants quickly and efficiently (Buhrmester et al., 2018; Stewart et al., 2017) and running behavioural studies online has become relative common (Grootswagers, 2020). Moreover, there has been also an increase in studies utilising so-called gamification approaches, where game-like elements (competition among online gamers, smooth visual feedback, etc.) are brought to online data collection to increase participant enjoyment and engagement with the research task (Nacke & Deterding, 2017).

Here we examine the issues relating to online *implementation* of studies via online platforms and the *recruitment* of participants via recruitment services, with a specific focus on auditory research. Implementation covers techniques, web protocols, online hosting, and tasks and controls related to collecting good quality data online. Recruitment of participants refers to the way participants are contacted, either via traditional recruitment methods (participant pools, social media, etc.) or through recruitment services (such as Prolific.co or Amazon's MTurk). Together, studies which are both implemented online and also utilise recruitment services are often called *crowdsourcing* studies (Stewart et al., 2017). Lastly, gamification, which is a related but different style of recruitment and implementation of online study, will be briefly presented as the approach shares many of the considerations all online studies in auditory domain.

In this paper, we first introduce the typical tasks utilised in online auditory research, and present some common online platforms for hosting and organising online experiments. We move on to summarise the main recruitment services, and then proceed to review the pros and cons of online studies, online studies with recruitment services, and finally cover online studies using a gamified approach. Overall, the elements of the online studies can be visualised as a process cycle (see Figure 4.1) where the new elements beyond the participants, researcher, and the experiment consist of an *online platform* that hosts and runs the experiment, and a *recruitment service* – or alternatively traditional recruitment that brings the participants to the experiment. As discussed further below, the design of the research can be pre-registered, and the data obtained from the experiment can be shared according to Open Data initiatives in behavioural sciences (Shrout & Rodgers, 2018) as any empirical research, but in the text we address the specific issues of transparency unique to online studies.

4.1.1 Types of online tasks in auditory research

To orient the reader to the challenges of online research, we first outline four types of tasks (rating, forced-choice, production, and stimulus manipulation tasks) reported in online studies in the auditory domain.[1]

In *rating tasks* participants are asked to make decisions based on subjective qualities of stimuli in a continuous manner. Examples range from the perception

[1] A pool of recent studies was defined by searching Google Scholar with the search terms "headphone check Prolific.co" and "headphone check MTurk", as headphone checks have become an indispensable part of online auditory studies. These keyword combinations generated a corpus of around 100 studies from which we have drawn the majority of our examples.

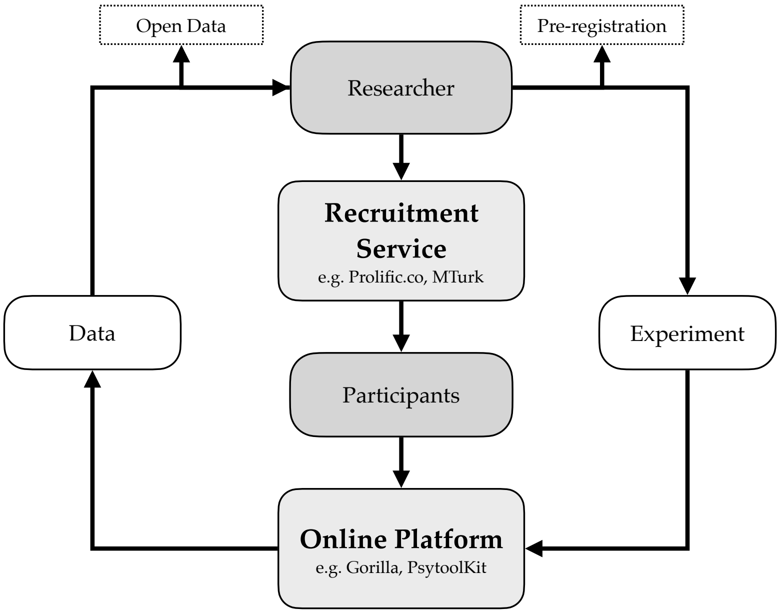


Figure 4.1: Diagram to illustrate the elements of an online experiment with a recruitment service and online platform.

of social traits from voices (Lavan, Mileva, Burton, et al., 2021; Lavan, Mileva, & McGettigan, 2021), to rating the consonance of intervals and chords (Lahdelma & Eerola, 2020), evaluating the "song-like-ness" of the speech-to-song illusion (Tierney et al., 2021), or rating the recognition of spoken words (Slote & Strand, 2016). In certain cases, these tasks can be implemented as indirect measures, where for instance, economic games can be used to assess the impact of perceived social traits from voices (Knight et al., 2021).

In *forced-choice tasks* the participant is exposed to different audio stimulus comparisons and asked to choose between the alternatives, for instance, classifying genre with different levels of noise (Kell et al., 2018), detecting a melody among variants (Woods & McDermott, 2018), or recognising syllables from audio-video presentation (Brown et al., 2018). Similarly, studies have examined how listeners learn to recognise people from voices (Lavan, Knight, Hazan, & McGettigan, 2019; Lavan, Knight, & McGettigan, 2019), speech intelligibility (Yoho et al., 2019) or how the perception of familiar and unfamiliar voices may differ (Kanber et al., 2020; Lavan, Kreitewolf, et al., 2021). Finally, studies have explored whether and how listeners can incorporate voices into their self-concept (Payne et al., 2020). One interesting variant of a discrimination task is the computerised adaptive beat alignment test (Harrison & Müllensiefen, 2018), where a participant is presented with a beep track and a musical track at the same time. The key variable is the temporal match between the beep track and the true beat locations, which are varied in an adaptive fashion during the experiment. In some forced choice tasks, the decision timing information can be used to gain insights about lexical or semantic decisions related to auditory signal (Armitage et al., 2021).

Production tasks can consist of singing, tapping, or altering parameters of sound/music. A good example is tapping with an external stimulus, which can be done in an online setting utilising the built-in microphone and speakers of a standard laptop computer to a high degree of temporal accuracy ($\tilde{2}$ ms latency and jitter) using REPP (Rhythm ExPeriment Platform) (Anglada-Tort et al., 2021). For a broad review of the timing capacities and accuracy of online platforms, see Bridges et al. (2020). Similarly, responses to tasks can be collected via singing in online experiments that also utilise the microphone input (Pfordresher & Demorest, 2020, 2021). Other examples of production tasks run in an online setting include a study where participants read sentences alone and in synchrony with another speaker in an online setting (Bradshaw & McGettigan, 2021), and a set-up in which adults had to try to learn novel words and were then required to verbally produce the words at test (James et al., 2020). Participants can also type their responses on a keyboard, which has been used to study speech perception (Guang et al., 2020; Heffner et al., 2017).

Stimulus manipulation tasks may employ sorting and arranging stimuli or controlling the parameters of the stimuli. For instance, Lavan and her colleagues (Lavan, Burston, & Garrido, 2019; Lavan, Burston, Ladwa, et al., 2019; Lavan et al., 2020; Njie et al., 2021) asked listeners to sort a number of voice recordings by perceived identities using a drag and drop interface. An example of auditory research where the participants controlled the creation of the stimuli according to specific criteria is the work by Harrison and others (Harrison et al., 2020). In their experiment, participants were asked to alter expressive parameters such as duration, pitch, and intensity of spoken prosody to create emotional expressions conveying one of three emotions. They also applied a similar methodology to enable participants to create pleasant musical chords by allowing them to modify the frequency of the intervals on a continuous domain. It is likely that there will be an increase in future in flexible and creative tasks that utilise visual layouts and new kinds of assessment and production tasks.

In summary, there is already considerable diversity in the approaches, techniques and methods utilised in online studies involving audio stimuli. However, most of the studies that we cite here are from the pre-pandemic era and it is reasonable to expect a significant growth in the number of online studies within the next years. Many labs and research groups have turned their attention to online opportunities because traditional lab experiments have been paused at the point of writing for more than

21 months during the periods of lockdowns or social distancing. It will be exciting to see how this current expansion of online studies will develop more robust and accurate ways of capturing our engagement with auditory stimuli and how these might shape empirical data collection practices – including lab-based behavioral experiments – in the near future.

4.1.2 Online testing platforms

There are numerous services that offer full online testing capabilities. Most of these have recently been reviewed in detail by Sauter and his colleagues (Sauter et al., 2020). The services can be divided into *integrated services*, which contain an experiment builder and host the live experiment data, such as *Gorilla.sc* (Anwyl-Irvine, Massonnié, et al., 2020), *Labvanced* (Finger et al., 2017), *Testable* (Rezlescu et al., 2020), *JATOS* (Lange et al., 2015) and *Pavlovia*[2]; and those that are mainly experiment builders but may also have capabilities to manage other aspects of the data such as *jsPsych*, *lab.js*, *OpenSesame*, *PsychoPy* (Peirce, 2007), *Qualtrics* and *PsyToolkit* (Stoet, 2010). We note here that "hosting the live data" refers to the actions whereby the service records, stores and possibly displays the choices made by the participants during the experiment, and does not refer to the long-term data sharing hosting solutions where the full data is deposited to an open access repository (e.g., to OSF, Github, etc.). It is also worth pointing out that most of the experiment builders (e.g., Psychopy, Gorilla, jsPsych, OpenSesame, Labvanced, and Psytoolkit) can run locally as well, which may well be desirable in situations requiring lab fa-

[2]https://pavlovia.org/

Table 4.1: Summary of the popular online testing platforms in auditory studies.

	Hosted	Fee	Example	Task
Qualtrics	Within	$300/month	Lahdelma & Eerola, 2020	Consonance rating
Gorilla	Within	$1/per participant	Lavan et al., 2019	Voice recognition
PsyToolkit	Within	Free	Armitage et al., 2021	Auditory & word priming
PsychoPy	Pavlovia	$145/month	Escudero et al., 2021	Auditory & visual priming

cilities, specialist interfaces (monitor, response device, or linking the responses with psychophysiology or neural responses), or in situations where the highest possible timing accuracy is needed. Here we focus on the platforms that have been featured in multiple auditory studies such as Gorilla, PsyToolkit, Qualtrics, and PsychoPy (see Table 4.1).

Qualtrics is a survey platform that supports JavaScript and html5 techniques. The platform interface for participants supports 75 languages, and provides good programmable control over the presentation orders (block and item randomisations) and allows the answers and the block choices to interact with responses. Qualtrics also supports links to recruitment services and have their own specific service called Qualtrics Panel. They are the most expensive in comparison to other services highlighted here, but universities often have institute- or department-wide subscription to the service. While Qualtrics handles complex survey and questionnaire designs and basic presentation of text, images, and limited audio and video, the specific needs of auditory research require an additional layer of coding in JavaScript. For instance, collecting timing responses in relation to audio or adjusting the audio presentation options requires controlling these parameters with JavaScript, which also requires specialized competence on how to integrate the code with the custom Qualtrics JavaScript Form Engine. For this reason, Qualtrics is not an optimal platform to develop complex, custom auditory experiments, although it would not be impossible.

Gorilla is an integrated platform that has an easy-to-use experiment design interface (task builder, questionnaire builder and experiment builder tools). It supports a wide variety of techniques, has tools for detecting bots, and supports recruitment services (Anwyl-Irvine, Massonnié, et al., 2020). It also achieved excellent precision (under 3 ms) in response time tasks although suffers from variability in the audiovisual synchrony in a direct comparison of the timing capacities of five online platforms (Bridges et al., 2020, more about these in section 2.1). Gorilla also offers a rich set of experiment and task libraries that can be used in building new experiments. Gorilla charges users for 'tokens', each enabling users to collect and download one data set. Users can either buy batches of tokens, or commit to lab/team/department subscriptions that may have different pricing strategies. Creating complex experiments and tasks is relatively easy with the task builder that relies on a graphical user interface. As such, Gorilla offers more flexibility in producing iterative, trial-based experiments than e.g. Qualtrics can easily provide. Additional functionality can be added to tasks by using snippets of JavaScript in the task builder. These snippets can be relatively easily implemented by those with only limited experience of JavaScript, while experts can set up entire tasks and experiments purely through code.

Psytoolkit is a free platform that contains experiment builders for experiments and surveys, and a large roster of example experiments and surveys. It also hosts the running of experiments and stores the data (Stoet, 2010). It is free for non-commercial use and geared for academic studies, especially for those using visual stimuli, but has good support for audio as well. It also supports integration of recruitment services in the data collection routines. There is no visual user interface and the surveys and experiments are programmed through a custom syntax. However, the examples and the help files are organised in a way that allow users to master the syntax after some period of study and consequently create sophisticated experiments with feedback options.

PsychoPy is another set of free tools for creating advanced behavioral experiments that can be run online (Gallant & Libben, 2019). It has a wide selection of techniques and experiment libraries. Online versions of studies are hosted at Pavlovia.org, but the integration of PsychoPy and Pavlovia is seamless. Whereas PsychoPy is free, hosting experiments in Pavlovia has a yearly cost ($2,053 or £1,500) or per participant cost (measured in credits). Pavlovia supports popular online techniques such as *jsPsych* and *lab.js*, which are high-level JavaScript libraries designed for creating behavioural studies with minimal programming experience. Like other online study platforms, it also supports integration of recruitment services in the data collection routines. The graphical user interface in the experiment builder makes this a suitable entry level tool, but competence in Python is useful to develop experiments that go beyond presenting auditory stimuli (e.g., to precisely control the timing of the visual and auditory stimuli).

4.1.3 Recruitment services (crowdsourcing)

When online studies are carried out with the help of recruitment services, we usually talk about crowdsourcing (Bhatti et al., 2020). This combination has the potential to be transformative for certain types of research questions and task within cognitive sciences (Stewart et al., 2017), and data sciences. In auditory studies, using recruitment services in conjunction with online studies is relatively common. With the rise of recruitment services such as Amazon's *Mechanical Turk*, *Prolific.co*, *Clickworker* or *Appen*, there are new opportunities which are used already by many scholars working on auditory issues. Table 4.2 pools together the summary details such as the size of the recruitment pool, the median hourly salary, citations, and citation frequency of audio-related studies for each of four popular services. We note, however, that the landscape of the available services may shift rapidly as there are new services such as *Sojump*, the Chinese recruitment service, and *CloudResearch* (formerly known as TurkPrime) and *Prime Panels* associated with CloudResearch, which provide additional ways of using Mechanical Turk with advanced customisation options.

Although there are papers already available that summarise several of these recruitment services, particularly Mechanical Turk (Chandler et al., 2019; Grootswagers, 2020), we nonetheless provide a short overview of the currently available popular recruitment services and their relation to auditory studies.

Amazon Mechanical Turk (hereafter MTurk) offers a large online workforce who can complete experiments and surveys at a very competitive price (Henrich et al., 2010). MTurk (and the alternative platforms discussed below) has proven to be more representative of the population than the usual lab or survey samples (Behrend et al., 2011): Participants using these services tend to be more reliable than typical online surveys or lab participants (Lakens, 2014), and even more attentive than lab participants (Hauser & Schwarz, 2016). Researchers can specify a number of criteria

Table 4.2: Summary of four popular recruitment services.

Name	Participants	Median Wage[a]	Auditory-related[b]	Citations
Mechanical Turk (US, 2005)	250,000[c]	$10.20	167	86,100
Appen[d] (AUS, 2007)	1,000,000[e]	$1.85	5	6,850
Prolific.co (UK, 2015)	153,308	$6.91	2	4,610
Qualtrics Panel (US, 2018)	Unknown	Unknown	37	1,980

[a]According to http://faircrowd.work although the median wages do not include the proportion of non-payment reported by participants, which are 60% for Mechanical Turk, 11% for Appen, 29% for Prolific.co according to http://faircrowd.work, see Section 3.1 for a full discussion.
[b]Web of Science search terms with auditory|audio|music|speech+recruitment service.
[c]Though 80K are estimated to be active each year (Stewart et al., 2015).
[d]Formerly known as Crowdflower.
[e]Unsubstantiated.

from the participants, including their past performance in tasks. Over the last ten years, large-scale studies with MTurk participants have successfully replicated classic studies in psychology (Berinsky et al., 2012), political science (Goodman et al., 2013), and economics (Mullinix et al., 2015). However, not all reports about MTurk are positive; several scholars have noted that the participants in MTurk are no longer naļve participants (Chandler et al., 2019). Overall, however, MTurk has been a popular recruitment service among researches, such that there are numerous studies involving auditory stimuli and participants recruited from MTurk (e.g., Aljanaki et al., 2017; Harrison et al., 2020; Howe & Lee, 2021; Schmidtke et al., 2018; Speck et al., 2011).

Prolific.co is another popular service that offers a large and flexible participant pool. Prolific offers a particular advantage to the academic community in that it is specifically designed as a platform for recruiting participants for research studies (Palan & Schitter, 2018). Prolific allows researchers to pre-screen participants for a range of variables that are pertinent to empirical auditory research, such as musical training, hearing loss, native language, handedness as well as linguistic and other demographic variables. Similarly to MTurk, Prolific allows researchers to include participants based on their past performance on the platform (approval rate, registration date, participation in different types of studies, number of completed studies, etc.). There are options to allow Prolific to offer representative samples for an additional fee or to balance certain demographic aspects of the samples. There have been a considerable number of studies involving auditory stimuli or music conducted with participants recruited from Prolific.co, including dozens of online experiments by the authors. Our studies have addressed consonance and dissonance of chords and interval using self-reports (e.g., Athanasopoulos et al., 2021; Lahdelma et al., 2021; Lahdelma & Eerola, 2020), auditory priming (Armitage et al., 2021; Armitage & Eerola, 2020; Lahdelma et al., 2020), perception of social traits from voices (Knight et al., 2021; Lavan, Mileva, Burton, et al., 2021; Lavan, Mileva, & McGettigan, 2021), or various voice identity perception tasks (Kanber et al., 2020; Lavan, Knight, Hazan, & McGettigan, 2019).

Appen is the third popular provider of recruitment services to researchers, which was known until 2019 as *Crowdflower*. Like MTurk, Appen provides a large and highly flexible pool of participants who are able to complete a range of tasks that include but are not limited to participating in academic research. Peer et al. (2017) contend that Prolific and Crowdflower provide higher quality data than MTurk, despite the latter's prominence as a research tool. While Appen is therefore another viable recruitment service that can be used for research, so far, there are to our knowledge no auditory or music studies that used Appen as a source of participants.

Qualtrics Panel provides customised recruitment services in the form of a panel of participants, which is negotiated with the company, and the pricing is related to the length of the survey and specificity of the sample required. Unfortunately the details of the panel participants (how many, what are the typical fees and so on) are not readily available, but several studies have compared Qualtrics Panel to other recruitment services in terms of the quality of the data and representativeness of the samples to nationally representative samples (Zack et al., 2019) in the US and in India (Boas et al., 2020). The results suggest that Qualtrics Panel may offer slightly more representative data than MTurk with specified criteria, and generally the notion of representative sample should be used with caution when recruiting through these services, especially outside the US. We have only identified one study relating to auditory stimuli and music that has used Qualtrics Panel as the recruitment service (Jakubowski et al., 2021).

As a basic visualisation of the increase in prevalence in crowdsourcing in auditory and music research over the last 15 years, Figure 4.2 shows the number of results returned by Google Scholar searches for the keywords "Music" or "Auditory" coupled with each of the recruitment services, "MTurk", "Prolific.co" and "Crowdflower" for each year from 2005 to 2020 (MTurk, Prolific.co and Crowdflower were launched in 2005, 2015 and 2007 respectively).

In addition to scholars capitalizing on recruitment services, there have been initiatives to build complete "virtual labs" or meta-services that function as a one-stop shop for online study needs. Such a service would provide the full framework which handles every aspect of the online research, from recruitment that taps into

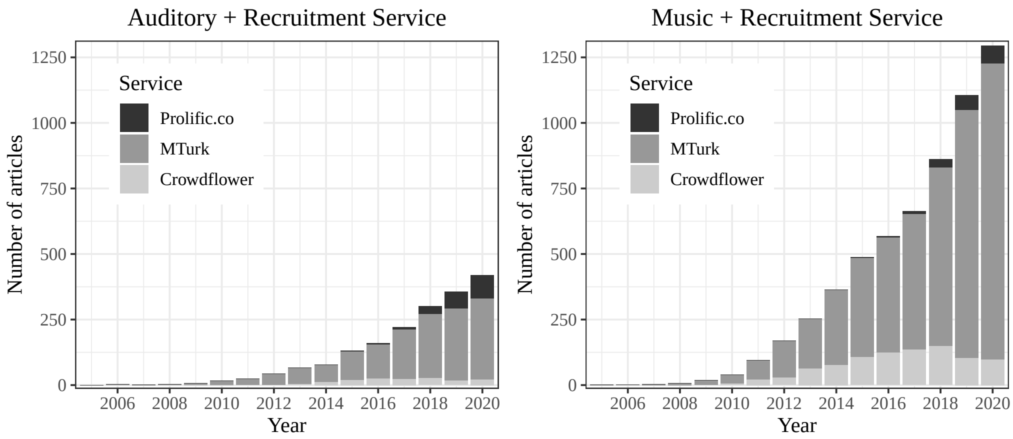


Figure 4.2: Frequency of the top three recruitment services in published studies since 2005 according Google Scholar results for "Auditory" or "Music" + recruitment service.

recruitment services (if needed), to implementing any online task and rendering the interface to the participant via a browser, preliminary quality control, payment and the bonus payment deliveries, and analysing and storing the data. Services that would support the full spectrum of online study processes such as EMPIRICA (Almaatouq et al., 2021), Dallinger [3], or PsyNet (Harrison & Jacoby, 2020) are all under development at the moment. While some of these might eventually become the convenient way of carrying out online studies, the extra advantage they offer is the capacity to support methods that require participants to interact with each other at a large scale. Studies involving iterated reproduction of rhythms (Jacoby & McDermott, 2017) or iterative altering of the properties of auditory stimuli (Harrison et al., 2020) are good examples of this approach that require integration over the distinct parts of the processes.

In the next section, we will review the pros and cons of the three approaches to online data collection compared to more traditional methodologies: online studies,

[3] https://github.com/Dallinger/Dallinger

online studies with recruitment services, and gamified studies.

4.2 Review of approaches

4.2.1 Online testing in auditory research pros and cons

Transparency. Online research fits well with reproducibility and open data initiatives although it is not inherently different from standard lab experiments. From the perspective of recruitment, platforms such as Prolific.co allow eligibility criteria to be applied automatically according to the registration information provided by participants. As a result, subsequent studies/researchers can recruit from a highly similar participant pool according to identical criteria. From the perspective of implementation, online tasks can be shared in their entirety (among collaborators for example via the "Library" space in Gorilla or sharing the experiment in Prolific.co or preferably publicly by allowing anyone to duplicate the study settings), including instructions, stimuli and response screens. This allows different teams of researchers to replicate studies with ease, or for data collection to be distributed between collaborators. Studies which are carried out online avoid lab-specific biases in terms of equipment, set-up, experimenter or delivery of instructions; and the ability to deliver the same web-hosted tasks both in the lab and in individual participants' homes allows for the robustness of findings to changes in environment to be assessed.

By sharing recruitment procedures and tasks, the research pipeline for online studies can be clearly defined and automated, and study pre-registration and open data release are also common strategies with this type of research, especially if

Dallinger or similar tools are utilised in deploying the study. A growing number crowdsourcing studies are pre-registered in OSF, although pre-registrations are still overall relatively rare in auditory cognition.

Democratisation of research. Online studies allow scholars with fewer resources – those without dedicated physical infrastructure (e.g., audio labs) or technical support services – to conduct studies by only paying the running costs of the experiment. Online studies also circumvent physical limitations beyond the global pandemic, such as smaller cities and towns having a smaller participant pools to recruit from, especially outside of university term time. The limitations of the size of the accessible participant pool may also impact piloting and estimating effect sizes adversely in these settings. Using recruitment services helps to avoid these issues and keeps the pace of research reasonable despite physical bottlenecks created by facilities or region.

Accessing specialist samples. Online research may offer the possibility of accessing specialist samples easily (Smith et al., 2015; Wilkerson et al., 2014). The speakers of specific languages (Turner et al., 2012), regional accents (Njie et al., 2021) or participants with amusia or tinnitus are examples of such special groups of interest to auditory research.

From this brief overview, it is clear that online studies do indeed provide benefits to the researcher on a number of levels. Additionally, the utility of online studies has come to the fore in allowing researchers to pursue experimental procedures where public health concerns or local restrictions would otherwise have made this impossible. It seems plausible that, even allowing for the rising trend demonstrated in Figure 4.1, we will see a further increase in the number of online studies that also employ recruitment services. However, the ease of data collection also presents several shortcomings, which are discussed next.

Sound delivery. In online studies, the researcher relies on the hardware and the software (mainly the browser) the participant already has access to on their own device. Participants can be asked to complete headphone checks such as those developed by Woods et al. (2017), which utilises phase-information to create differences in dynamics that are easy to discern with headphones but not with external speakers such as those in laptops, or Milne et al. (2020), which is based on the Huggins pitch test; code and stimuli for both tests are freely available online. Such tests allow researchers to screen out those that do not use headphones[4]. At the time of writing, twelve peer-reviewed studies that use Woods et al.'s 2017 headphone check are available and for which the pass-fail rates are reported. The mean and median failure rates are 16.5% and 17.2% respectively (range: 0% - 40%). One study (Guang et al., 2020) utilises the test by Milne et al., and found a failure rate of 49%. There is also a new headphone check available relying on beating interference to verify the participant's hardware capacity to present stimuli dichotically (Pankovski, 2021).

Even with such headphone checks, the quality of the delivery of the stimuli as well as the quietness of the environment is not under the experimenter's direct control. This puts serious limitations on more psycho-acoustically demanding studies (aiming to establish detection thresholds, just noticeable differences [JNDs], etc.). Moreover, there are ethical issues related to the potential exposure of participants

[4]Anecdotally, it is possible to pass the Wood et al. (2017) test reliably via certain laptop speakers

to unexpectedly or unintentionally loud sounds in online studies. Due to this, we recommend that studies should always start by presenting some kind of representative "calibration sound", with participants instructed to start with their volume turned down and then adjust it to a comfortable level. Experimental stimuli should then never be louder than the calibration sound, and we stress the importance of adjusting the sounds carefully (starting quietly and turning the volume up, rather than vice versa). However, despite such calibration checks and clearly worded instructions, the experimenter cannot fully ensure that the participant's equipment and setup would not lead uncomfortable listening experiences for the participant.

Aside from those specialist auditory studies that are best carried out in well sound-proofed settings, there can be some merit to the argument that if a phenomenon can be captured in the diverse settings such as those offered by participants' typical headphones and audio devices, the phenomenon is probably a robust one. However, imperfect playback devices may also bias the results in specific ways instead of simply increasing noise. For example, if the effect of interest relies on low frequency information in the auditory signal, this is typically poorly represented by the average headphones (Olive et al., 2018). Even though typical listeners are unable to detect the quality differences between budget and high-end headphones (O'Brien & Schmidt, 2020), results from online studies in such cases may differ from results obtained in more controlled settings. In cases where the sound quality may become an issue for adequately measuring effects of interest, we would recommend at least running validation studies in the lab.

Timing accuracy. Online experiment platforms such as PsyToolkit, Gorilla, and

PsychoPy that run in most current browsers all demonstrate good capacity to bring timing experiments, such as reaction time measures in response to visual and auditory stimuli, to an acceptable level of precision; they usually demonstrate inter-trial variability of 5-10 ms, as compared with lab-based software, which can achieve inter-trial variability under 1 ms (Bridges et al., 2020). For a large-scale comparison of these qualities for lab-based and web-based software across multiple operating systems and browsers, see Bridges et al. (2020) and Anwyl-Irvine, Dalmaijer, et al. (2020). In a nutshell, the two evaluation studies demonstrate that web-based solutions provide adequate timing for most cases unless the absolute response times are needed between the individuals. This measure is negatively impacted by different participants providing their responses on different operating systems and browsers. This, however, is not normally required when the comparisons are made within the same participant as is often the case in experimental research. An important caveat to the timing accuracy of the online data collection is that significantly better response timing accuracy is obtained using external response devices (i.e., high-performance button box) than by using the standard keyboard (Bridges et al., 2020). For this reason, the poorer accuracy overall reported by Anwyl-Irvine, Dalmaijer, et al. (2020) is closer to the reality of online research as the participants will not have high-performance button boxes installed on their USB ports.

Answer format. Many standard answer formats are supported across the various online testing platforms, such as different types of questionnaire responses, forced-choice responses, free text responses and ranked responses among many others. However, not all types of answer formats that might be constructed in a lab are feasible in an online experiment. Because of the dependency on participants' home setups, it is not always possible to use unusual interfaces which require complex mouse operations, or calibrated monitors or other external devices, e.g., any kind of neuroimaging or physiological recording. Thus, for the most part, experimental procedures must be limited to using standard mouse operations and keyboard responses. Despite these limitations, recent research has demonstrated the feasibility of production tasks, such as tapping (Anglada-Tort et al., 2021) or singing (Pfordresher & Demorest, 2021), or adjusting sliders to create sounds (Harrison et al., 2020) in studies of auditory cognition. Similarly, even in the absence of support of a specific answer format on online testing platforms, alternative ways of implementing tasks can be found: For example, in the absence of a readily available drag-and-drop interface at the time, we have run a sorting study online, asking our participants to download a PowerPoint slide on which the to-be-sorted stimuli were embedded. Participants were then able to sort the stimuli within PowerPoint, save the sorted slide and upload to a file transfer website from which we were able to retrieve the data (Lavan, Burston, & Garrido, 2019; Njie et al., 2021).

Lack of visual oversight of participants Although many lab experiments do not require visual connection during the experiment between the participants and the experimenter, lack of any visual – or auditory – cues during the experiment can amplify problems that sometimes occur in lab experiments such as participants attempting to engage in social media, text message or calls, or becoming perplexed by the experiment instructions or getting stuck at some point. In our view, issues such as these are less likely to occur when participants attend labs in person and interact with experimenters; or in case of confusion with respect to tasks, they can be resolved with timely interactions. Overall, these issues are the crucial part of the quality control that we articulate in more detail in section 4.3.5.

Copyright and other restrictions. Much of the stimuli used in auditory and cognition research might be copyright free sound files created for the experimental needs, but in cases where existing commercial music or audio excerpts are used, the uses of the copyright material needs to comply with regional law governing fair use and digital copies of copyrighted materials.

4.2.2 Online testing using recruitment platforms pros and cons

Diversity. Ethnocentrism – i.e. focusing too heavily on one particular subset of the human population – is one of the main criticisms levelled at psychology in recent years (Rad et al., 2018). By turning to recruitment platforms, we may be able to avoid some of the aspects of so-called WEIRD samples (White, Educated, Industrialized, Rich, and Democratic) (Casler et al., 2013; Henrich et al., 2010) as the participants in these services have more diverse backgrounds (Sheehan, 2018) than typical participant pools. Specifically, Casler, Bickel and Hackett found that MTurk samples are more diverse than traditional volunteer (recruited via social media) samples in terms of both their socio-economic and ethnic backgrounds 2013, also supported by Chandler et al. (2019). Goodman et al. (2013) found a greater degree of linguistic diversity in MTurk samples than in a typical community sample. However, this is not to say that the diversity provided by MTurk is still far away from

US population: most Mturk participants are young and older people are underpresented in the participant pool. MTurk participants tend also be more liberal, and have higher education qualifications when compared to the US population (Casey et al., 2017; Levay et al., 2016). However, some authors point out that the greater degree of variation in the participant pool can also act as a limitation (Feitosa et al., 2015) in some contexts, such as when certain instruments have not been validated in a particular language.

For music and auditory studies, it is worth highlighting that the criticism of WEIRD is extremely relevant as music is highly culturally dependent and notions and preferences about music do vary considerably even within a country depending on sociodemographic background and education. A particular aspect of diversity is cross-cultural research, which normally requires extraordinary connections and resources. Online studies with recruitment services offer a possibility to tackle some, albeit limited cross-cultural research (Cuccolo et al., 2021). Most of the recruitment services allow researchers to define participant recruitment by location, country, and native language at least. This allows for comparisons that relate to geographical location and language of the participants, although it has to be kept in mind that the popularity of the services is not well spread beyond the countries of their originators, and that the people who work in these services are very much reliant on internet and may have fairly Western standards in many of their values (Pollet & Saxton, 2019).

In our experiments using participants obtained from recruitment services such as Profilic.co, we have observed that samples are more representative in terms of their age, gender, and nationality than the average student population, although they do nonetheless deviate from the society at large. The employment details are stable across many experiments and an overall fairly even gender distribution can be readily achieved through pre-screening. The age distribution usually shows that the bulk of the participants are between 25 and 42, and about 20% of the participants are students, which is an improvement from lab studies but it is still clear that students are over-represented in comparison to national statistics (3.5% in the UK[5]).

Affordability. Online studies may provide a cost-effective way of collecting data in auditory sciences. Provided that the quality assurances can be met, the price of data may be cheaper than in lab studies when you factor in the costs involved (researcher time, research assistants, facilities and participant fees). For instance, we carried out a direct economic comparison of lab and online data utilising recruitment services with respect to a specific study (Armitage & Eerola, 2020). When including the cost of both participant fees and payment to a research assistant, the cost of the lab data was more than double that of the online data. However, the direct lab costs may often be lower if they include free labour available as part of the operation (research assistants working for course credit etc.) although this cost could be assumed to included elsewhere (such as training of the assistants and general costs of facilities and services). As an example, a typical online dataset requiring 10-minutes of participant time using Prolific.co and a sample of 40 participants would cost \$106.40 (\$12 for minimum wage $\times$ 1/6 h $\times$ 40 participants $\times$ 1.34 service fee) excluding any piloting. Importantly, there were no significant differences in the

[5]https://www.hesa.ac.uk/news/17-01-2019/sb252-higher-education-student-statistics/numbers

attrition rates, distributions of the data, or effect sizes between the samples obtained via recruitment services and lab (Armitage & Eerola, 2020). However, it should be noted that Buhrmester et al. (2018) advocates for careful investigation of attrition rates, although this is a factor in all internet-mediated research and not a problem that is unique to studies using recruitment services.

Speed and statistical power. Online studies with recruitment services allow for rapid data collection. Studies that would take several weeks to run in a lab can be completed in the course of a day via recruitment services depending on the type of task and specialist expertise needed. Clearly, fast data collection offers direct benefits in terms of time saved. The speed of data collection also provides more nuanced benefits beyond speed per se. As data collection is efficient in terms of time, it allows for research to focus on testing specific hypotheses. Funding allowing, hypotheses can therefore be revised and retested at a rate that was not possible previously, allowing for an incremental but thorough advancement of understanding. Also, online studies with recruitment services allow for studies to be appropriately powered, not over or under-powered, if the power analysis is made in advance and with conservative reading of similar studies (Brysbaert, 2019). We therefore argue that the improvements to research are most pronounced if the advantage conferred by the speed and easiness of the data collection is tempered with carefully planning and appropriate assessment of the needs.

Data Quality. Traditional volunteer web studies are often subject to significant data wastage or poor data quality (Hochheimer et al., 2019). However, there is research to indicate that the quality of data collected via recruitment services is comparable to lab data and superior to other web data and subject to less participant attrition (Armitage & Eerola, 2020; Hauser & Schwarz, 2016; Kees et al., 2017), although contrasting views also exist (Chmielewski & Kucker, 2020; Grootswagers, 2020). Indeed, we have found that participant attrition and prevalence of outliers in samples obtained through recruitment services is almost identical to lab data and superior to traditional online data collected via convenience sampling (Armitage & Eerola, 2020). Data quality is of course related to the task, instructions, and the quality controls implemented in the study, which we cover in the later (section 4.3.5).

Scalability. Online studies with recruitment services allow the possibility of automating the data collection at a high level, which allows easy transportability, replicability and the possibility to alter the method, concepts, stimuli or a measure by a simple option or even run several variant experiments in parallel. This is also possible at some level in traditional lab experiments, but with considerable more effort and customisation. The real benefit of the scalability comes from experiments where participants' responses are taken as input for other participants such as in iterated rhythm production tasks (Jacoby & McDermott, 2017) or in Gibbs sampling with humans (Harrison et al., 2020). Scalability usually requires that the researchers utilise an automation service such as Dallinger or Pushkin (Hartshorne et al., 2019), which allow a high level of abstraction and the automatation of all practicalities (recruitment, running the experiment, paying participants, and managing data).

Collecting data with recruitment services also brings several possible shortcomings, presented next.

Quality control. The reliability of online studies utilising recruitment services has been explored with several well-known psychometric instruments (personality etc.) and these indices (obtained with test-retest evaluations and calculating Cronbach alphas) are typically at the same level as in lab experiments or in surveys (Buhrmester et al., 2018). However, in addition to quality control covered in Section 4.3.5, there have been reports of a number of instances of fraud and other quality issues, especially in MTurk. Some of these have been attributed to fraudulent respondents who typically use VPS/VPN (virtual private servers and networks) to hide their identity (at least nationality or the specific IP address to allow multiple submission, etc.) from the service (Kennedy et al., 2020). For this reason, it has been suggested that data from respondents connecting to recruitment services via VPS/VPN should be discarded. In a similar vein, Kan and Drummey (2018) found that a significant minority of participants were willing to carry out experiments despite not meeting inclusion criteria. Due to these concerns, Kan and Drummey have proposed some mitigating measures, such as using the demographic pre-screening offered by the recruitment services where possible rather than by stating inclusion criteria when advertising the study to participants.

In addition to fraudulent participants, there have been reports of bots being present in MTurk (Chmielewski & Kucker, 2020; Kennedy et al., 2020; Moss & Litman, 2018). Depending on their sophistication, bots can be identified by them failing even simple quality control checks and should provide low-quality data on even the simplest task. Similarly, bots could be, for example, identified by anomalous response time behaviour and by coming from few specific geolocations, although much of the prevention has to take place at the recruitment services, since they are able to see the full pattern of data and also enforce policies on how new accounts are verified and approved[6]. Analyses carried out by the recruitment services themselves suggest that the problems were not created by bots, but by a small number of foreign workers related to a few server farms. As such, the quality issues should, therefore not be a problem as long as sufficient data quality controls and checks are in place within an experiment.

Longitudinal or interconnected studies. Longitudinal studies can be challenging in many of the recruitment services. For instance MTurk does not natively support longitudinal studies or studies where participants need to be pre-screened by criteria that are not included in MTurk's standard demographic profile. There are ways to work around this limitation (Stoycheff, 2016), however, and services such as *CloudResearch* facilitate implementing various operations with the same participants over time that allow running longitudinal studies and they also provide additional quality control, see Chandler et al. (2019). Follow-up studies are supported in Prolific.co and Gorilla, where it is possible to reinvite participants who completed previous parts of studies, or to exclude participants who have completed previous related studies.

Sampling issues. Although participants from recruitment services might offer a more diverse samples than those typically obtained in most of the lab studies using well-educated undergraduate students, participants from recruitment services tend to be younger and less likely to be fully employed than national averages (Mellis

[6] Prolific.co's anti-bot measures, see https://blog.prolific.co/bots-and-data-quality-on-crowdsourcing-platforms/ and https://gorilla.sc/online-experiments-and-bots-what-can-be-done/ and analysis of these events at CloudResearch https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/

& Bickel, 2020). As such, recruitment services do not really provide a population sample but a convenience sample. This is unlikely to be a large problem for auditory research. If prevalence estimates are needed, or research questions require a representative sample, the recruitment service needs to be used in a specific way to capture the characteristics of a representative sample. Some of these services, Prolific.co and Qualtrics Panel, offer a separate service where the researcher can request a representative or a stratified sample from the population for an additional fee.

Research practices. An obstacle for research utilising online data collection with recruitment services may be that peer-reviewers may not yet be familiar with using online testing and recruitment services. This can lead to queries voicing concerns about experimental control, sample characteristics and the overall data quality. Depending on the task, it may, of course, be valid to ask for a lab validation. However, validations have been run already for many routine experimental paradigms and tasks, limiting the usefulness of further lab validations. In our experience, reviewers' concerns can be addressed by having included quality control measures, such as headphones screening and attention checks. Given the increased use of online studies with or without recruitment services, we expect that this type of data collection will soon be considered as valid and standard, provided that appropriate quality controls are in place.

Another issue related to samples is that the ease and the speed of research may also tempt scholars to adopt questionable principles such as p-hacking, HARKing, or other problematic practices that go against the traditional use of statistical thresholds (Wicherts et al., 2016). There are solutions to some of these issues such as performing sequential analyses during the data collection (Lakens, 2014) to avoid p-hacking, or pre-registering the study intentions, outcome measures, sample size, and inclusion/exclusion criteria. The use of preprint servers (e.g., PsyArXiv[7] or arXiv[8]) to post all studies soon after they are concluded will guard against the danger of only reporting those iterations of the study that delivered results under the conventional statistical thresholds. But overall, these problems exist in any empirical research and require commitments to research integrity rather than special measures to control online studies using recruitment services.

4.2.3 Online testing using a gamification approach – pros and cons

Gamified online studies utilise "game-design elements in any non-game system context to increase users intrinsic and extrinsic motivation, help them process information, help them to better achieve goals, and/or change their behavior." (Treiblmaier et al., 2018, p. 134). Gamification has been occasionally utilised in music and audio-related online studies; there are several online games addressing rhythm (Bellec et al., 2013; Duffy & Pearce, 2018), an online game for collecting music similarity data (Wolff et al., 2015), and an online game for detecting hooks in music (Burgoyne et al., 2013) as well as projects going on about musicality and other topics[9]. The benefits of gamified data collection can be substantial (Honing, 2021). Online games

[7] https://psyarxiv.com
[8] https://arxiv.org
[9] e.g., https://www.themusiclab.org/

can potentially lead to a very large number of participants without paying them anything. They also may help to spread the word about topic and increase engagement and the impact of research. However, not all studies can be gamified and the very nature of creating a game suitable for anyone may work against research goals or prevent researchers from imposing necessary controls or from collecting crucial background information. Gamified data collection tends to require very bespoke development and typically requires considerable investment in app development or web technologies, although some of the online testing platforms are now starting to include "game-builders" in their services[10]. Generally gamification studies are difficult to replicate as the public interest will wane after the initial wave of curiosity. All critical issues of implementation that are relevant for online studies in the auditory domain in general are relevant to gamified studies as well, but we see gamification as a special approach that may offer a unique combination of engagement that brings in limited data from a diverse yet a large sample of interested people.

4.3 Key commitments when utilising recruitment services in online auditory research

Conducting research with recruitment services rather than traditional online surveys or lab experiments raises ethical questions that we want to address next. We address issues that may be of concern to individual Institution Review Boards as well as considering good practice more broadly. Whilst some of the points we raise could

[10]See https://gorilla.sc/product/gorilla-game-builder/

be discussed in the context of *all* empirical research, many are unique to online studies using recruitment services and auditory research.

4.3.1 Ethical considerations

The first assumption is that all empirical research including both online studies in general and those studies which use recruitment services – should be subject to the local ethics policies, which usually implement the national research integrity and ethics regulations.

The use of recruitment platforms – mainly MTurk – has received negative press in recent years both within and outside the academic community (see, for instance, Semuels, 2018). The concerns raised have suggested shortcomings in the behaviour of individual researchers and in the governance of the platforms. The focus of these concerns is frequently financial, with very low rates of remuneration being reported as common and quite possibly the norm (Hitlin, 2016). Despite the median hourly wage of $10.20 reported in Table 4.2, Hara et al. (2018) have calculated that when non-payment and returned tasks are taken into account, the median hourly wage on MTurk is in the range of $1.77 - $2.11. Non-payment refers to the proportion of participants in the service that report being not paid at least once for their work. These numbers are surprisingly large: 60% for Mechanical Turk, 11% for Appen, and 29% for Prolific.co according to http://faircrowd.work. It is also worth pointing out that services such as faircrowd.work are able to pool together experiences from hundreds of participants in these services. In addition to fair payment issues, there are additional concerns such as anonymity, misrepresentation of task duration or complexity by researchers, or unfair rejection of work (Salehi et al., 2015).

As outlined in section 1.3, Prolific.co was launched as a recruitment platform specifically for academic research. Alongside the benefits to researchers that we have presented already, Prolific.co has inbuilt safeguards to ensure fair treatment of participants. Payment is at a recommended minimum rate of $10.26 (£7.50) per hour with a hard minimum of $6.84 (£5) per hour. If researchers reject (i.e., do not pay) a participant, then they must provide the participant with a reason, and there is clear guidance to researchers as to when a participant's submission can and can not be rejected, with suggested alternatives such as allowing the participant to redo the task or offering partial payment. There is also guidance for participants on how to appeal rejection decisions. Prolific's supporting documentation for researchers provides reminders for researchers about these policies which aim to ensure fairness.

4.3.2 Reporting commitments

As for studies run in the laboratory, researchers should commit to reporting all technical solutions and decisions for studies using online testing and recruitment in a format that enables the replication of these studies (e.g. by specifying the recruitment filter(s), headphone check(s) and its pass-rate, stimulus preparation, visual materials, inclusion/exclusion criteria). It is also important to report the date range of the data collection as the construct of interest may change over time, or we may learn that the recruitment service's participant pool was compromised with "bots", fraudulent participants or a surge of newcomers. There is also the possibility to share the experiment fully via the online platform and also share the recruitment service details within the recruitment tool, thus allowing others to capitalise on the exact same tasks, protocols, stimuli, instruction and sampling criteria. Nothing prevents researchers from releasing the stimulus, design, and the de-individualized data[11] in an Open Access repository[12] and also include the experiment scripts (and analysis scripts, for that matter). Prolific.co promotes transparency in reporting practices, advocating for use of mechanisms such as pre-registration or registered reports. Gorilla hosts Open Materials to share the protocols, tasks, and questionnaires[13]. Likewise Buhrmester et al. (2018) suggests that researchers report in detail on the use of the recruitment platform, for instance any restrictions on participants' experience, attrition rates, rates of payment, and so on. We suggest that these details would be well-suited to be formal reporting under online experiment protocols, similar to guidelines for reporting experimental protocols in life sciences (Giraldo et al., 2018) or bio- and nanosciences (Faria et al., 2018).

4.3.3 Financial commitments

Participants from recruitment services often have a dual role as both 'workers' and 'participants', and a proportion of participants fall below the Federal Poverty Line (Ipsen et al., 2021), yet several participants report hourly remuneration less than US Federal minimum wage (Hitlin, 2016). Thus, it is the researchers' responsibility to ensure that rates of remuneration are fair, and do not simply reflect the minimum remuneration possible. To align with lab-based studies, we encourage researchers to

[11] Data from recruitment services is not fully anonymous since the participant recruitment service IDs and IP addresses are often logged into the data

[12] E.g., https://osf.io, https://github.com, or https://dataverse.org)

[13] https://app.gorilla.sc/open-materials

commit to pay the respondents *at least the minimum wage*. In case of estimation of the duration of the task being overall too short in comparison to the actual time spent on the task, it is recommended – and is relatively easy in most recruitment systems – to increase the participant fees – or pay bonuses – to reflect the actual time spent on the task. An interesting dilemma is the currency and wage differences between the countries. Is the compensation tied to the country of the participant or the researcher? It might be safe to err on the side of higher pay, but that will also set up pressures for participants from low income countries to take part with VPS/VPN posing as coming from another country to make significantly higher earnings.

4.3.4 Fair use of recruitment services

In response to the financial and ethical issues raised about the use of recruitment services by academic researchers, several universities have enacted policies on the use of recruitment platforms in research, and the platforms themselves provide (often non-binding) guidelines on fair and professional treatment of participants.

1. *Transparency.* Ensure that tasks are transparent in terms of ownership: There is a clear ownership that can be traced back to the PIs and ethics approvals, much the same as there would be for in-lab experiments or more traditional internet-mediated research. For participants, it should be very clear and transparent what the remuneration is for a task, the amount of time it will take, timescale for payment, and whom to contact if questions arise.

2. *Valuing participants.* Ensure participants are aware of the value of their contribution to the research. Offer a debrief at the closing stages, a lay summary,

suggestions for further reading, and thank the participants along with the payment.

3. *Professional standards.* Recruitment services handle the payments in different ways, but one of the main issues for participants is that they are paid fairly and promptly. Some of the services auto-approve participant payments and partial payments (due to failing attention checks or otherwise not completing) are organised in various ways. Also make sure that all responses to participant queries are prompt and professional and possibly utilise similar standards even if handled by separate researchers and research assistants.

4. *Rejection policy.* The policy of rejecting participant contributions should be clear. Rejecting participants' work may have implications in some recruitment services for payment for both the present task and the participants ability to access other tasks. There are cases where rejection is likely such as when the participant fails to pass attention or headphone checks. In these circumstances participants can be reimbursed only partially (e.g., if they contributed a few minutes of their time before failing the headphone check) or not at all (e.g., if they fail attention checks over a certain tolerance such as 20% of the attention checks). If it becomes known that a participant has started the task, but not completed it (e.g. from timestamps or direct contact from the participant), it is possible to pay a pro-rata equivalent for the time spent on the task. For data that are not usable, it may be possible to give participants the opportunity to redo the task. It should be noted that some platforms, such as Prolific.co reserve the right to overturn rejection decisions.

4.3.5 Quality control

Over the past 5 years, the authors have utilised Gorilla, Qualtrics and PsyToolkit for online studies and when the need has arisen, turned mostly to Prolific.co or occasionally to MTurk as the recruitment service. Most of the studies we have carried out online been relatively straightforward data collection exercises without complex elements such as follow-ups or pre-screening, although these have been implemented in some cases. Based on these experiences and following the ongoing scholarly discussion about online studies and use of recruitment services, we want to highlight the topic of quality control, which was not explored in detail during the earlier discussion of pros and cons.

The topic of quality control has received considerable attention with regard to online studies, since the assumption is that the participants working remotely will have more distractions, a wider range of backgrounds and life situations, less uniform expectations of what to do in the studies, or even fraudulent motivations to participate in studies, all of which could lead to unwanted variability in the responses. However, Jennifer Rodd (2019) argues that many of the quality control checks designed for online studies should actually also be implemented in lab studies. We agree that quality control should be an inherent part of any data collection, not just limited to online studies. One such operation is to design the experiments in labs and online contexts to be within-subject designs, which mitigates the differences between the different set ups (equipment, volume, etc.) as well as some of the individual differences.

As the data collection environment is not under the experimenters' control in an online study, it is important to provide quality checks that ascertain whether the participant is paying attention to the task at hand and understands the instructions properly. Here we divide these checks into generic attention, technical, consistency, expertise, and honesty checks:

1. *Generic attention checks* such as *Instructional Manipulation Checks* (IMCs) can be used. In this the respondent is shown the following text:

   > "You should not answer this question if you read it; it is to check your attention: (1) Strongly Disagree; (2) Disagree; (3) Don't Disagree/Don't Agree; (4) Agree; (5) Strongly Agree"

   Past research had demonstrated that 16% to 18% of respondents fail IMCs, although this rate is not higher than in lab studies (Paas et al., 2018). Variant attention check can be tasks that resemble captchas (Completely Automated Public Turing test to tell Computers and Humans Apart) where participants are asked to pick a colour, word, or an image have been utilised as attention checks and to eliminate bots.

2. *Domain specific attention checks*, that is, attention checks that rely on auditory information can be used. Such a task can ask participants to "type the two digits your hear in a speech excerpt into the box below", e.g. Sauter et al. (2020) or be presented as a variant of a captcha relying on timbre, pitch height or another auditory property of interest.

3. *Technical checks* typically relate to audio quality such as checks for headphones and general ability to discriminate volume or pitch differences (Milne et al.,

2020; Woods et al., 2017). For production studies such as capturing tapping or singing, there are usually initial checks to test the recording and timing capacities of the computer (Anglada-Tort et al., 2021). One can also use a "honeypot" checks which targets only bots by implementing two forms on top of each other, where a human participant only sees one, but any automated script will see and offer responses to both (Downs et al., 2010).

4. *Repeated items* is another way of measuring attention and allows the researcher to analyze the possible inconsistencies in the responses.

5. *Expertise checks* relates to self-disclosed expertise, which is again not unique to the online studies as it applies to all studies, but the accessibility of these experiments and the potential payments received from these may encourage prospective participants to mislead the experimenter about their background or expertise. To ensure that participants meet the expertise criteria, it is recommended to test the specific expertise rather than rely on self-reported expertise. To give an example, for expertise about a specific musical genre (for instance, Hindustani classical music), one can devise short, timed statements that require the specific expertise to answer correctly. Or preferably, the questions can be in the form of audio examples ("Is this sound example in North Hindustani (a) Dhrupad, (b) Khyal, (c) Ghazal or (d) Thumri style? (please choose one)").

6. *Honesty checks* can be implemented after the main task: Participants can be asked whether they truly fit the recruitment criteria, while being assured that their answer will in no way affect their payment. It is known across multiple

studies that a noteworthy proportion (3-28%) of participants are dishonest about their qualifications (MacInnis et al., 2020).

The quality control in lab studies is normally implemented post-experiment, with unreliable responses/participants discarded based on pre-defined criteria (such as intersubject reliability, response speed, or another task-dependent criterion). It would be possible to implement the control protocols within the online experiment and eliminate the inattentive or noisy respondents during the experiment, but in our experience and according to the principles outlined in the payment policies above, we have deemed it safer to assess the quality of the responses and participants after the experiment.

4.4 Conclusions

Overall, online studies and studies with recruitment services can be set up relatively easily compared to traditional lab studies, and can often be carried out more quickly. However, this comes with a trade-off in terms of implementation (control of environmental conditions, attention, and audio setup) and recruitment. It is possible to mitigate the implementation issues to some extent, for instance, by use of headphone checks or attention checks embedded within the experiment, but online studies will always be carried out on equipment of variable quality in situations that are varied across the participants. Online studies that draw the participants from recruitment services seem to be less subject to the degradation of quality between lab and online data compared to volunteer web samples recruited for instance via social media. Indeed, we have not found a significant difference in the quality of the data between lab studies and online studies using recruitment services, either in terms of participant attrition or in the distribution of the data itself (which has been the case with online volunteer samples).

Online studies may not be inherently more transparent than traditional lab studies, but many of the open science principles such as pre-registration, sample size determination, and replication are at least somewhateasier to implement than in traditional studies. The transparency is also promoted by some recruitment services (Prolific.co[14]). They also allow the eligibility criteria to be applied automatically and in principle facilitate the recruitment of highly similar set of participants in subsequent studies. It is also possible to share entire experiments within the online experiment system, making at least direct replications straightforward. Finally, online studies can also avoid specific biases that labs may have in terms of facilities, equipment, experimenter, or instructions.

Overall, online studies with or without recruitment services do not remove the necessity for lab studies in auditory research, but they do allow for good quality data to be collected outside of a lab. Recruitment services can be seen to offer several advantages over convenience samples. As well as offering opportunities for data collection when access to labs is restricted, for instance during the Covid-19 pandemic, online studies offer benefits in their own right. As scholars in music cognition and other auditory disciplines grow more accustomed to the benefits and challenges of online studies, they are bound to become more frequent in coming years. We hope that our reflections and summaries above are helpful to researchers

[14]https://researcher-help.prolific.co/hc/en-gb/categories/360000850653-Prolific-s-Best-Practice-Guide

embarking on their online studies and promote good practice in terms of research transparency, quality control and ethics.

4.4.1 Acknowledgments

We thank Peter C. Harrison for insightful comments about gamification approach and for detailing many of the issues of scalability that come with the use of systems such as Dallinger and Pushkin.

CHAPTER 5

Introduction to the empirical chapters

Abstract

Chapters 6, 7, 9 & 10 are made up of discrete published articles (Armitage et al., 2021; Armitage & Eerola, 2020, 2022; Armitage et al., 2023); at the time of writing, Chapter 8 is under review. The purpose of the present chapter is to orient the reader to the overarching narrative of the chapters, introducing the main theoretical considerations and summarising a proposal for a model of affective priming based on the findings of the empirical chapters. In addition, this chapter details the common aspects of the procedure and analysis strategies whilst directing the reader to key differences in the procedure, stimuli and experimental manipulations. Finally, this chapter outlines the procedure for developing the music stimuli used for the

experiments in Chapters 6, 7 and 8.

5.1 Introduction

The empirical chapters that follow divide naturally into three sections. Firstly, Chapter 6 verifies the validity of collecting reaction time data (RT) online by comparing RT distributions for participants in three different conditions (lab data collection, online collection via social media, online collection via crowdsourcing services). Secondly, Chapters 7 and 8 consider the cognitive mechanisms implicated in affective priming. The final pair of empirical chapters considers priming from the perspective of stimulus features, and then probes whether expertise can alter responses to stimulus features. Chapter 9 considers how the psychoacoustic properties of the stimuli can contribute to priming. Chapter 10, considers how pyschoacoustics and expertise in a specific musical culture, Lithuanian Sutartinės contribute to automatic responses (indexed by affective priming) versus self-report responses to chord stimuli. Table 5.1 outlines key aspects of each chapter.

Chapter	Article	Prime Stimuli	SOA	Participants
6	Armitage & Eerola (2020)	Affective Music	450 ms	lab vs social media vs Prolific
7	Armitage & Eerola (2022)	Affective Music	450 ms	non-specific, recruited via Prolific
8	Armitage & Eerola (under review)	Affective Music; Environmental Sounds	450 ms	High vs Low anxious
9	Armitage et al. (2021)	Dyads (piano; Shepard tone)	200 ms	Non-specific, recruited via Prolific
10	Armitage et al. (2023)	Dyads (vocal timbre)	200 ms	Sutartinės singers, Lithuanian controls, Western musicans, Western controls

Table 5.1: Prime type, SOA and Participant Pool for Chapters 6 - 10

Broadly speaking, the methods for each chapter follow the same outline, employing a primed word classification task. In all cases the primes are auditory stimuli. There is, therefore, a high degree of overlap in the procedures in each of these chapters. The prime (i.e., music stimuli) are common to Chapters 6, 7 & 8, and the target words are common to all the empirical chapters. The development and choice of stimuli is detailed below in Section 5.3 below.

5.1.1 Theoretical accounts of affective priming

There are several theoretical accounts of affective priming (for discussion of these accounts see Herring et al., 2013; Rohr & Wentura, 2022). Here, the two main explanations are briefly outlined: spreading activation and Stroop-like interference. In addition, the evidence accumulation model of binary choice tasks (of which affective priming is one) is described.

Spreading activation

Fazio et al. (1986) put forward an early account of affective priming based on activations in semantic memory. Semantic memory is modelled as a network, with each node representing a semantic concept. Exposure to a stimulus partially activates related nodes in the memory network. Consequently, responses to subsequent related (i.e., congruent) stimuli are facilitated because the relevant node in the memory network is already partially activated.

Stroop-like interference

An alternative explanation for priming is based on Stroop-like interference. This account is based on *response competition.* Participants evaluate the target as positive or negative and respond accordingly. However, it is assumed that participants also automatically evaluate the prime stimulus. In the case of incongruent conditions, the conflict between the tendency to respond to the prime and the response to the target results in targets typically being classified more slowly and less accurately in incongruent conditions compared to target conditions.

Binary classification as evidence accumulation

Although not developed specifically for affective priming, Roger Ratcliff's Ratcliff, 1978 evidence accumulation conception of binary choice phenomena also provides a useful framework for considering affective priming. Under this model, humans determine whether to classify a stimulus as positive or negative once enough evidence has been accumulated to cross a decision threshold.

5.1.2 Overview of findings and proposed model of affective priming

Overall, the empirical chapters confirm the consensus view outlined in Chapter 3 that music primes are capable of generating affective priming. In particular, both dyads and pre-composed music clips are capable of acting as primes. However, in the case of dyads, priming was only present when the negative prime had a high degree of acoustic roughness.

Chapter 7 provides at least partial support for the task-specific attention allocation argument proposed by Spruyt et al. (2009). That is, the dimension of affect that is activated by the prime depends on where a listener focuses there attention. However, it should be noted that valence seems to be transferred in both valence and arousal priming. This chapter also contained a semantic priming task, a lexical decision task. Neither valence or arousal was detected in this task. The absence of priming effects in this task suggests that automatic evaluation of a concept taskes place only if it is task-relevant. When words are categorised as positive or negative, the valence of the music inluences the speed of valence classification (and similarly for arousal priming), but when the decision is 'word vs not word' the affective content of the music does not influence the speed of the evaluation of the target. Thus this result seems to suggest that, without attending to the affective features of the target, the prime does not pre-activate similarly-valenced concepts in the memory network. In turn, this supports a Stroop-like account of affective priming.

In Chapter 8, congruency effects were present when both music and environmental sounds were used as primes. However, the predicted modulation of the congruency effects in the High Trait Anxiety group was only present when environmental sounds were used as primes. In particular, responses to positive stimuli were disrupted by the presence of a negative prime.

Chapter 9 is consistent with Chapters 7 and 8 in finding priming only when one of the intervals is high in acoustic roughness. There was no group difference in congruency effects – i.e., there was no evidence that expertise or enculturation does not influence automatic (that is, affective priming) responses to acoustically rough intervals.

The results of the empirical chapters, particularly Chapters 7 and 8, indicate that Stroop-like interference perhaps provide a better account of affective priming than spreading activation models. The next section explores in more detail a proposal for a theoretical account of priming that integrates several of the themes explored in the empirical chapters.

Proposed model of affective priming

How can Stroop-like interference, top-down vs bottom-up attention be integrated into one one model, that also accounts for the critical finding from Chapter 3 that affective priming is sensitive to SOA? One possibility is that, under the evidence accumulation model, a participant deploys top-down attentional resources to undertake the word classification task. A participant determines whether a target word is positive or negative based on the accumulated evidence. However, this task can be disrupted by the information contained in the signal from the prime. In the case of dyads, it seems that the only circumstance under which a negative prime contains enough information to disrupt the classification of a positive target is if the dyad is acoustically rough – acoustic roughness is thought by some authors to convey threat-related information. Likewise this would also explain the results of a related study Lahdelma, Armitage, and Eerola (2022) which found that tetrads (four-note chords) induced priming, but that dyads did not. One explanation is that (particularly dissonant) tetrads are higher in acoustic roughness than dyads; an alternative is that a tetrad is made up of three intervals. Both explanations are consistent with the notion that it is the amount of information that is conveyed that is responsible for disrupting the top-down task. However, it could also be argued that if the roughness account is correct, then it is the *urgency* of the information rather than the *amount* of information that is critical. Further evidence for information accumulation is that in the case of the music primes, negative low arousal stimuli function as effective primes, even at the slightly longer SOA of 450 ms. It is also plausible that some combination of amount and urgency of information is responsible for the action of the primes. For instance, in the case of negative, low arousal primes at SOA = 450 ms, there is a combination of negative valenced musical cues (e.g., dynamics, timbre, tonality); in the case of short exposure to a rough dyad, the relative paucity of information is less important as the high degree of roughness orients the listener to the negative information via a bottom-up attentional process. Indeed, the combination of threat-related urgency and rich information may account for the ability of negative environmental sounds to activate anxiety related biases in affective priming, whereas music – which is rich in cues, though perhaps denotes less ecologically valid threat – does not activate these biases. Future research should probe more specifically the relationship between measures of musical information (e.g., event density, combinations of affective cues such as mode, timbre, tempo, harmony, chord numerosity), SOA, and affective priming.

5.2 Methodological details

Next, this chapter provides an outline of the common details of the procedure and analysis strategy for the priming experiments. The chapter concludes by outlining the development of the music stimuli which were used in Chapters 6, 7, & 8, and the target words that were used in all empirical chapters.

5.2.1 Outline procedure

Participants were asked to classify target words as positive or negative by pressing the 'm' and 'z' keys respectively on the computer keyboard. Participants were instructed to classify the words as quickly and accurately as possible. Target words were immediately preceded by auditory primes. However, there is variation in SOA and choice of primes. Chapters that employ ecologically valid music primes use a longer (450 ms) SOA compared to chapters that use chords, where the SOA is 250 ms in line with previous chord priming literature (Costa, 2013; Lahdelma, Armitage, & Eerola, 2022; Sollberger et al., 2003; Steinbeis & Koelsch, 2008, 2011). Furthermore, choice of particpant is a key (quasi) experimental manipulation in Chapters 8 (participants high in trait anxiety vs participants low in trait anxiety) and 10 (Sutartinės singers vs Western Musicians vs Lithuanian Controls vs Western controls).

5.2.2 Analysis of reaction time data

In Chapters 7, 8, 9 and 10, RTs are fitted to a Generalised Linear Mixed Model (GLMM) following the Gamma family with an identity link function (Lo & Andrews, 2015). Data was subject to pre-treatment such that:

- incorrect answers were deleted prior to analysis
- timeouts were treated as incorrect answers and deleted from the data set
- participants whose accuracy rate was below 75% were excluded from the analysis
- lower outliers – i.e., items with a reaction time of less than 250 ms were deleted from the analysis
- upper outliers[1] were identified by fitted each particpant's RTs to an ExGaus-

[1]A set of pilot results was analysed under several outlier protocols; choice of outlier protocol did not substantially effect the results of null-hypothesis significance tests.

sian distribution (Ratcliff, 1993). RTs that exceded the 95th percentile of the distribution were deleted.

.

GLMMs were then subject to a type III Anova that reported Wald χ^2 values to test if the factors in the models were significant.

5.3 Stimulus development and selection

The development of the music stimuli used in chapters 6, 7 and 8 is described in detail here; details of the chord stimuli used in chapters 9 and 10 are described in the materials section of those chapters.

Many music priming studies rely on simple chordal stimuli (e.g. Costa, Steinbeis & Koelsch, Bakker and Martin), whereas others, such as Goerlich (2011; 2012) and Lense (2014), make use of pre-existing clips of music designed to represent or induce specific emotions (see Chapter 3 for more detail). The present thesis uses both kinds of primes to probe different aspects of affective priming. Here, we report the procedure and analysis used to develop the music stimuli used as primes in Chapters 6, 7, & 8. Emotions are typically characterised using one of two established approaches: basic emotions, which categorise emotions as happy, sad, disgust, surprise, anger, or as some combination of those emotions. In this model, emotions are categorised by their valence (i.e., is the emotion positive or negative?) and by their arousal level (i.e., is the emotion high arousal or low arousal). Thus emotions are considered to span a two-dimensional space as demonstrated in Figure 5.1. Figure 5.1 maps individual emotions to their position in 2-dimensional valence-

Figure 5.1: Dimensional model of emotion (Russell, 1980)

arousal space. Note that, in more recent formulations of dimensional models, some authors further subdivide arousal into tension arousal and activation arousal. The research reported in this thesis favours a dimensional model of emotion over a basic emotion model. Thus, the objective of this section is to develop a set of 1000 ms music stimuli that capture via emotional cues (timbre/tempo/tonality etc) the four affective valence-arousal quadrants shown in Figure 5.1.

5.3.1 Method

Participants

42 respondents (17 female, Mean age = 37.3, SD = 18.4) completed the online pre-rating task. Respondents were recruited online via Reddit and SurveyTandem. Ethical approval was given by the Music Department Ethics Committee, Durham

University and participants gave informed consent via an online checkbox.

Materials and Stimuli

An initial set of candidate stimuli was found by carrying out a web search for musical stimuli that were used for emotion induction procedures. Two collections of stimuli were identified: Eerola and Vuoskoski (2011) and Västfjäll (2001). Following initial inspection of the stimuli, 32 stimuli were chosen to represent the four valence-arousal conditions. Next, we selected an 11 - 12s segment from each stimulus that was chosen from the set suggested by Vastfjäll which in the opinion of the researcher captured the properties of the valence-arousal quadrant in question; those taken from Eerola & Vuoskoski were already trimmed to between 11 000 and 14 000 ms. Manipulation of the audio stimuli took place in Audacity; the questionnaire was coded in PsyToolkit (Stoet 2010; Stoet, 2017)

Procedure

Next, we carried out an online survey in which participants were asked to rate stimuli on a scale of 1 - 7 for valence and arousal using an on screen slider. Rather than using the terms valence or arousal, the constructs were accessed using the proxies negative/sad/angry - positive/happy/cheerful and relaxing/calm - exciting/arousing. Participants were able to replay the stimuli as many times as they liked before rating the stimuli. The online interface is shown below in Figure5.2.

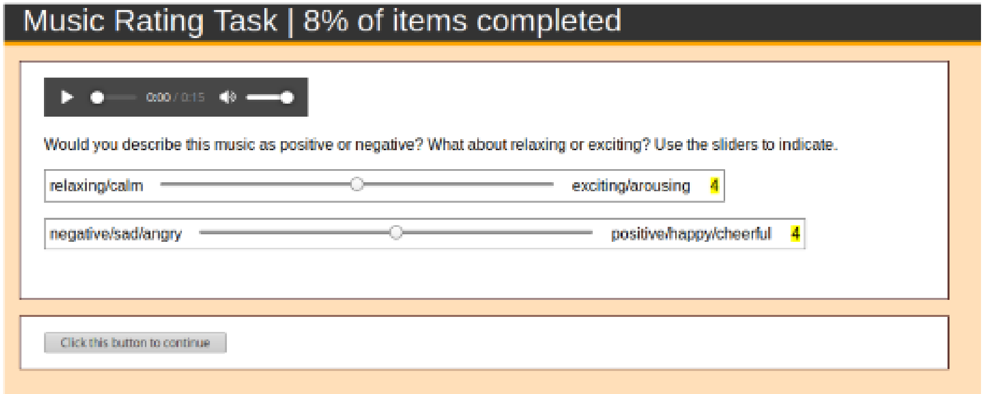


Figure 5.2: PsyToolkit screen for rating task

Results

There was moderately good agreement between raters, with ICC = .66 with a 95% CI of [.58, .74], $F\ (63, 2583) = 86$, $p < .001$. Higher values of ICC are associated with rating tasks where there is less subjective response and especially where raters have undergone some form of training, thus the reported result seems strong in this context, indicating a high degree of accordance between the raters. Mean valence and arousal pairs were calculated for each stimulus and are presented in Table 5.2. The ratings are also demonstrated graphically in Figure 5.3. (The coding $(\bar{x}, \bar{y}) \mapsto (\bar{x} - 4, \bar{y} - 4)$ is used to aid visual interpretation.) At first sight, the stimuli seem to be distributed in a satisfactory manner, with a clear cluster of stimuli in each quadrant. Only one stimulus, excerpt 31, was rated as not being in the expected quadrant.

Stimulus selection

Sixteen stimuli were selected - four from each quadrant to represent each of positive valence - high arousal, positive valence - low arousal, negative valence - high arousal,

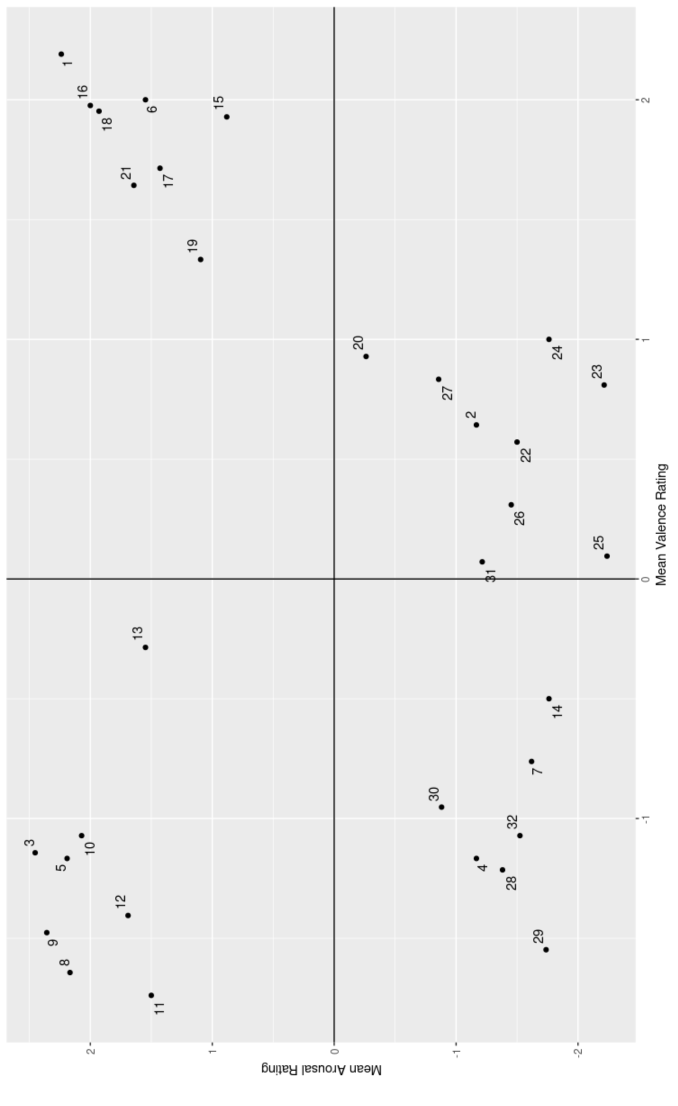


Figure 5.3: Valence-arousal ratings for 32 candidate music stimuli

negative valence low arousal. (The four quadrants will hereafter be referred to as PH, PL, NH, NL). We selected the 4 stimuli in each quadrant with the greatest Euclidean distance from the origin. The four stimuli per quadrant with the maximum distance are indicated by a † in Table 5.2:

Excerpt	Valence	Arousal	distance	Track	Quadrant
1	2.19	2.24	3.13	Sousa - Stars and Stripes Forever†	PH
2	0.64	-1.17	1.33	Holst - The Planets: Venus†	PL
3	-1.14	2.45	2.71	Holst - The Planets: Mars†	NH
4	-1.17	-1.17	1.65	Albinoni - Adagio†	NL
5	-1.17	2.19	2.48	Stravinsky - Rite of Spring	NH
6	2	1.55	2.53	Tim Weisberg - The Good Life†	PH
7	-0.76	-1.62	1.79	Grieg - Aases Todd†	NL
8	-1.64	2.17	2.72	001 Lethal Weapon 3 (Tr 8)†	NH
9	-1.48	2.36	2.78	002 The Rainmaker (Tr 7)†	NH
10	-1.07	2.07	2.33	003 The Allen Trilogy Tr (9)	NH
11	-1.74	1.5	2.3	004 Cape Fear (Tr 1) †	NH
12	-1.4	1.69	2.2	005 The Fifth Element (Tr 19)	NH
13	-0.29	1.55	1.57	006 Crouching Tiger, Hidden Dragon (Tr 8)	NH
14	-0.5	-1.76	1.83	039 Shakespeare in Love (Tr 3)	NL
15	1.93	0.88	2.12	021 The Rainmaker (Tr 3)	PH
16	1.98	2	2.81	024 Man of Galilee (Tr 2) †	PH
17	1.71	1.43	2.23	071 The Untouchables (Tr 6)	PH
18	1.95	1.93	2.74	075 Batman (Tr 18)†	PH
19	1.33	1.1	1.73	073 Shine (Tr 5)	PH
20	0.93	-0.26	0.96	074 Shine (Tr 15)	PH
21	1.64	1.64	2.32	072 Man of Galilee (Tr 2)	PH
22	0.57	-1.5	1.61	041 Shine (Tr 10)†	PL
23	0.81	-2.21	2.36	042 Pride & Prejudice (Tr 1)†	PL
24	1	-1.76	2.03	043 Dances with Wolves (Tr 4)†	PL
25	0.1	-2.24	2.24	044 Pride & Prejudice (Tr 12)	PL
26	0.31	-1.45	1.48	047 Oliver Twist (Tr 8)	PL
27	0.83	-0.86	1.2	049 Juha (Tr 2)	PL
28	-1.21	-1.38	1.84	031 English Patient (Tr 18)†	NL
29	-1.55	-1.74	2.33	033 Portrait of a lady (Tr 9) †	NL
30	-0.95	-0.88	1.3	034 Big Fish (Tr 15)	NL
31	0.07	-1.21	1.22	037 Batman (Tr 5)	NL
32	-1.07	-1.52	1.86	038 Dracula (Tr 7)	NL

Table 5.2: Valence-arousal ratings, distance from origin, track name and valance-arousal quadrant of candidate stimuli; where numbers are given in the track name, these refer to the track's listing in Eerola & Eerola (2011). Tracks indicated with a † were selected for use as experimental stimuli.

Next, the valence-arousal ratings for the stimuli were subject to various pairwise comparisons with Bonferroni correction to ensure that the four groups of stimuli were significantly different from each other in terms of valence and arousal. For ease of presentation, the results of the comparisons are given in Table 5.3.

Stimuli	Valence Comparison	Arousal Comparison
NH vs NL	.66	<.001
PH vs PL	<.0001	<.001
PH vs NH	<.001	.15
PL vs NL	<.001	1.00
PH vs NL	<.001	<.001
NH vs PL	<.001	<.001

Table 5.3: *p*-values for pairwise comparisons by stimulus group with Bonferroni correction

As predicted we saw significant differences in valence ratings between the positive and negative valence groups, but the negative-negative comparison did not yield any significant differences, nor did the positive-positive comparison. When comparing arousal, as predicted we saw significant differences in arousal ratings between the low and high arousal groups, and there was no significant difference in arousal between the NH-PH or the NL-PL stimuli. However, we did see a significant difference in valence ratings between the positive and the positive-low stimuli. In particular, we saw that positive-high stimuli were rated as being significantly more positive than positive-low stimuli This difference is discussed in more depth below.

Next, in order to shorten the sixteen shortlisted stimuli to a length suitable for use in a priming experiment, a panel of 6 expert listeners drawn from the fields of musicology, music and emotion, and performance identified one 1000 ms clip per excerpt that best captured the valence - arousal quadrant that the clip was chosen to represent. The duration of 1000 ms was chosen following the line of reasoning pursued by Scherer and Larson (2011) that music stimuli evolve in time and so the longer extract length compared to the chord studies (where the typical duration is 800 ms) was chosen. The clips had a mean length of 1015 ms (range 1000 to

1045). The variation in length was permitted to preserve the musical integrity of the stimuli. Finally, the 1000 ms excerpts were rated by 32 non-expert listeners for valence and arousal. The results of this additional rating task are reported below.

Valence-arousal ratings for one-second extracts from music stimuli

35 participants, who did not take part in the priming task, rated the 1s stimuli outlined in Table 7.2.1 on both valence and arousal dimensions on a 9-point Likert scale. The mean rating values are shown in Table 5.4. For brevity, stimuli names are coded by valence-arousal quadrant rather than by the tile of the musical extract.

Stimulus	Valence	Arousal
PH1	6.1	7.0
PH2	7.3	6.3
PH3	5.9	7.0
PH4	6.2	7.4
PL1	5.4	3.2
PL2	6.1	3.0
PL3	6.7	2.7
PL4	6.3	3.5
NH1	4.9	7.5
NH2	3.5	7.7
NH3	2.9	7.3
NH4	2.4	7.5
NL1	4.8	3.0
NL2	5.0	2.9
NL3	4.2	2.9
NL4	4.2	3.2

Table 5.4: Mean (SD) valence and arousal ratings for music stimuli

The valence-arousal ratings for the stimuli are represented graphically in Figure 5.4

Finally, valence and arousal ratings for the 1000 ms stimuli were subject to Bonferroni-corrected pairwise comparisons

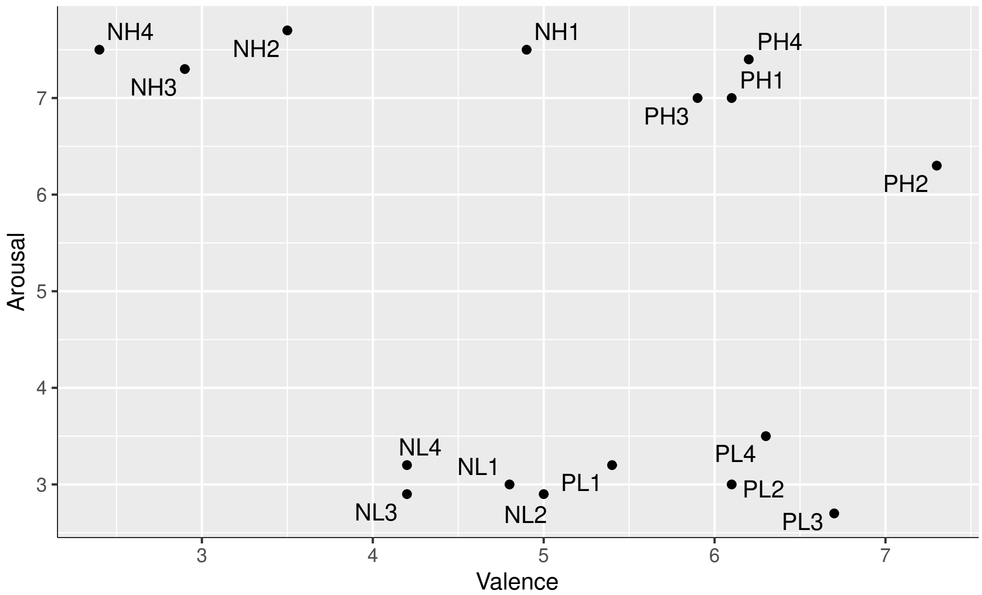


Figure 5.4: Valence-Arousal ratings for 1s extracts from music stimuli

Stimuli	Valence Comparison	Arousal Comparison
NH vs NL	0.27	$<$.001
PH vs PL	1.00	$<$.001
PH vs NH	$<$.001	.09
PL vs NL	0.05	1.00
PH vs NL	0.02	$<$.001
NH vs PL	0.001	$<$.001

Table 5.5: p-values for pairwise comparisons by stimulus group (1000 ms extracts) with Bonferroni correction

Discussion

Previous research has identified a potential confound between arousal and valence (Scherer & Larsen, 2011). Globally, the present data set has a very low degree of correlation between valence and arousal, and prima facie seems to avoid this confound. However, the difference in valence rating between the positive-high and positive-low groups presents a potential challenge. However, the 1000 ms extracts from the stimuli were not subject to this potential confound in either the valence or arousal dimensions.

5.3.2 Target words

In all instances, the target words were drawn from Warriner et al. (2013), a corpus of roughly 14000 English words rated on a 1 – 9 likert scale for valence, arousal and dominance by 1827 participants. Sixteen words were chosen as the the target stimulus set, four each to represent the quadrants positive valence-high arousal, positive valence-low arousal, negative valence-high arousal, and negative valence-low arousal. The words were *climax, lively, excite, snazzy*[2], *gentle rest, comfy, relax, rabid, hijack, arrest, fatal, saggy, flaccid, dismal, morgue.* The words consisted of four to six letters and one or two syllables. Table 5.6 reports Warriner et al.'s valence-arousal ratings for the target words.

Table 5.6: Target words for affective evaluative tasks grouped by valence-arousal (Valence-Arousal ratings given in brackets)

Positive-High	Positive-Low	Negative-High	Negative-Low
Climax (7.5, 6.8)	Gentle (7.4,3.2)	Rabid (3.0, 6.6)	Saggy (2.6, 3.0)
Lively (7.1,6.1)	Rest (7.9,2.2)	Hijack (1.8,6.1)	Flaccid (3.55,3.14)
Excite (7.8,6.6)	Comfy(7.2,2.9)	Arrest (2.3,6.7)	Dismal (2.6,3.3)
Snazzy (6.5,5.4)	Relax (7.8,2.4)	Fatal (2,6.8)	Morgue (1.8,3.5)

5.3.3 Data availability

Raw data, PsyToolkit scripts, R scripts and stimuli are available at https://osf.io/3tbpq/?view_only=18a60b72c94545f28f2ef83b0908a981

[2] *Lover* and *payday* were used in pilot testing. *Lover* was associated with very fast RTs; two participants reported finding *payday* ambiguous.

CHAPTER 6

Reaction time data in music cognition: Comparison of pilot data from lab, crowdsourced and convenience web samples[1]

6.1 Introduction

Reaction time (RT) methods have been a mainstay of research in cognitive psychology for over a century. RT methods have been applied in domains as diverse as visual perception (e.g. Ando et al., 2002), personality traits (e.g. Robinson & Tamir, 2005) and social psychology (e.g. Wang et al., 2017). In music cognition, RT methods have been used as an indirect measure of several phenomena such as harmonic expectation (Bharucha & Stoeckig, 1986), melodic expectation (Aarden, 2003) cross

[1]Published as Armitage, J., & Eerola, T. (2020). Reaction time data in music cognition: Comparison of pilot data from lab, crowdsourced, and convenience web samples. *Frontiers in Psychology*, 2883. https://doi.org/10.3389/fpsyg.2019.02883

modal priming (Goerlich et al., 2012), absolute pitch (Bermudez & Zatorre, 2009; Miyazaki, 1989) and emotional responses (Bishop et al., 2009).

Traditionally, reaction time data has been collected in a lab. However, recent years have seen the development of software capable of collecting accurate response time data online, for instance PsyToolkit (Stoet, 2010, 2017), PsychoPy (Peirce et al., 2019), Gorilla (Anwyl-Irvine, Massonnié, et al., 2020) and Qualtrics' QRTEngine (Barnhoorn et al., 2015) amongst others. In the early days of web-based reaction time studies, there was considerable scepticism about the viability of RT data collected online. Despite the prevalence of software specifically designed to collect reaction time data online, and the increasing incidence of Web-based data collection, there remains a degree of caution around online reaction time studies. However, recent research (Barnhoorn et al., 2015; de Leeuw & Motz, 2016; Hilbig, 2016) suggests that online reaction time data is perhaps more trustworthy than was previously thought, but these studies have not yet involved music as stimuli.

Alongside the developments in software, recruitment of participants in online studies has been made easier by the prevalence of social media and crowdsourcing platforms such as Amazon's MTurk service and Prolific. Not surprisingly, the use of crowdsourced samples by researchers is growing rapidly (Stewart et al., 2017).

However, to the authors' knowledge (with the exception of de Leeuw and Motz) the comparisons of laboratory and online RT data have focused on descriptive measures of the RT distributions, and relatively little attention has been paid to the agreement between the RT distributions as a whole. Moreover, none of these studies considers phenomena associated with music cognition. Given the widespread use of RT methods in music cognition and the growth of crowdsourcing as a recruitment tool, the authors consider there to be a need to test the viability of online RT collection specifically in the case of music cognition.

The present data report offers the results of a response time task completed in three different contexts — in a standard lab setting (Lab"), online recruited via 'traditional' online techniques (Web) and crowdsourced vis Prolific.ac (CS). Below, we present summary data for the three data sets before testing the comparability of the three data sets on an item-by-item basis.

6.2 Data collection

6.2.1 Reaction time task and stimuli

Data was collected using PsyToolkit (Stoet, 2010, 2017) for the lab and both online samples. PsyToolkit offers a choice of either a local installation in Linux or a browser-based version that can be used to collect data online. The PsyToolkit script used for the Lab, Web and CS data collection was identical in all three cases. Participants completed an affective priming task in which they heard a short (approximately 1000 ms) extract of music (.wav files in the Lab sample; .mp3 in the Web and CS samples) before being presented with a visual target word. Participants had to classify each word as positive or negative as quickly and accurately as possible. There were eight music primes and eight target words resulting in $8 \times 8 = 64$ prime-target pairs. The music primes, which were drawn from Eerola and Vuoskoski, 2011 and Västfjäll, 2001 were controlled for valence and arousal, as were the eight target words, which were taken from Warriner et al., 2013. There were two music primes in each valence-arousal condition: 2×positive-high, 2×positive-low, 2×negative-high and 2×negative-low. The target words followed the same valence-arousal distribution. Following the Lab data collection, it was found that one of the target words, *Lover*, was associated with significantly faster reaction times than the other words and was subsequently replaced with *Payday*. Both *Lover* and *Payday* have been excluded from the analysis below, leaving fifty-six prime-target pairs. Details of how the music clips were chosen and rated and more precise information regarding the procedure are included as supplementary material.

6.2.2 Lab study

Participants were all right-handed (Hardie & Wright, 2014; Kalyanshetti & Vastrad, 2013) with normal or corrected to normal vision and hearing; all were native English speakers and received £5 to complete the present study and a related study. Data were collected during June 2018. The experimental setup comprised a Lenovo laptop running Linux (Xubuntu 18.04) and PsyToolkit version 2.4.1.(Stoet, 2010). Including form-filling, the section of the experimental sessions relating to this task took around ten minutes.

6.2.3 Convenience Web sample

The materials and procedure mirrored the lab data collection as closely as possible, with the exception of the replacement of *Lover* with *Payday* as described above. Additionally, audio files were converted to .mp3 format. Data was collected using the web-based version of PsyToolkit (version 2.5.2) (Stoet, 2017) during July 2018. The script used was identical to the script used for the Lab experiment. PsyToolkit allows researchers to restrict which type of devices are used to carry out online experiments, so we excluded tablets and mobile phones in order to maintain as much similarity with the lab setup as possible.

Participants were recruited online via Reddit, SurveyTandem (a survey exchange website where researchers complete each others studies in exchange for points; when researchers have amassed enough points, their studies are made available for other researchers to complete) and student email distribution lists at the University of Durham. Participants received no direct payment for participating, but had the option of entering a draw for a £25 Amazon voucher. As with the Lab sample, the inclusion criteria were right-handed native speakers of English with normal or corrected to normal vision and hearing.

6.2.4 Crowdsourced sample

Participants were recruited via Prolific (www.prolific.co) and received a payment of £0.75. Owing to the similarity to a previous study that recruited via Prolific, participants from this previous study were excluded from taking part. Participants were prescreened to be right-handed, native speakers of English. The stimuli and procedure were identical to those used for the convenience web sample. The PsyToolkit version was updated to 2.5.4: the differences between the versions focused on the user (i.e. researcher) interface and did not impact RT collection. Data collection took place during July 2019.

6.3 Comparison of data sets

6.3.1 Data pre-treatment

Participants whose accuracy rate fell below 75% were excluded from the analysis. This resulted in no deletions from the Lab data, but six participants in the Web sample and two in the CS sample failed to reach the required accuracy threshold. For the remaining participants, timeouts and response times shorter than 250ms were excluded from the analysis, as is common practice (e.g. Duckworth et al., 2002). To exclude upper outliers, individual participants' response time distributions were fitted with an exponentially modified Gaussian (ExGaussian) distribution (Ratcliff, 1993). Responses above the 95th percentile of each ExGaussian distribution were removed from the data set. Removal of timeouts and outliers accounted for the deletion of 5.7%, 6.3% and 6.1% of responses from the Lab, Web and CS data sets respectively.

6.3.2 Comparison of summary data

Following deletions, there were thirty-two participants (mean age = 24.0, 19 male) in the Lab sample, thirty-three (mean age = 25.1, 13 male) in the Web sample and thirty-four (mean age = 32.7, 8 male) in the CS sample.

The three data sets are compared in accuracy, attrition rate, mean and variance of response time in the fifty-six prime-target conditions. Summary data is contained in Table 6.1.

The mean (SD) percentage error rates for the Lab, Web and CS samples were 3.67

(0.188), 3.64 (0.187) and 3.30 (0.179) respectively. Linear mixed effexts modelling suggested that there was no significant differences in accuracy rates between the Lab, Web and CS samples, $F(2, 110) = 0.32, p = .728$. A repeated measures ANOVA was carried out to compare the mean response time for each target-prime pair. The test proved non-significant, $F(2, 54) = 1.883, p = .16$.

Similarly, a repeated measures ANOVA was carried out to compare the variances in response times for each target-prime pair. There was a highly significant difference in variances in response times between the Lab (mean Variance = 26396) and Web (mean Variance = 16227) or CS (mean Variance = 18742) samples, $F(2, 110) = 26.22, p < .0001$. Contrary to expectations, post hoc testing indicated that variance in the Lab RTs was greater than the variance in Web or CS RTs.

Table 6.1: Summary Statistics for RT Distributions

Method	Error Rate (%)	Mean	Variance	Median	IQR	Timeouts (%)
Lab	3.67	580.34	15959.44	538	161	5.75
Web	3.64	587.352	16278.84	564	133	6.28
CS	3.30	587.98	18741.63	564	140	6.14
Combined	3.53	585.30	20267.40	557	148	6.06

6.3.3 Comparison of RT distributions

In addition to the comparison of the summary data carried out above, we also carried out overall and per-item comparison of the RT distributions in the three data sets. Figure 6.1a shows the overall cumulative RT distributions of the three data sets.

Following the procedure set out by Voss et al., 2013 to compare response time distributions for binary choice data, incorrect responses were allocated a negative response time (for instance an incorrect answer with a response time of 450 ms was coded as -450). Next, Kolmogorov-Smirnov (KS) tests were carried out to compare the Web Convenience vs Lab response time distributions for all 56 prime-target pairs. Eight out of the fifty-six results returned significant results, suggesting that, in these instances, the Web Convenience sample and the Lab sample could not be thought of as representing the same underlying RT distribution. A binomial test was carried out with $n = 56$, $r = 8$ and $p = .05$ to determine whether 8 instances of disagreement between the Lab and Web Convenience samples is more than can be expected by chance.This test returned a significant ($p = .006$) result suggesting that the overall distributions of RTs for the prime-target pairs of the Web Convenience and Lab samples cannot be considered equivalent.

The same procedure was carried out to compare the RT distributions for the Lab and CS data sets. As before, we carried out KS testing to determine the goodness of fit between the two RT distributions for each prime-target pair. Two conditions yielded significant results. Binomial testing ($n = 56, r = 2, p = .05$) confirmed that 2 out of 56 conditions is below the threshold for significance ($p = 1$), suggesting a strong agreement between the Lab and CS data sets.

6.3.4 Presence of hypothesized effects

Although reporting the results of the priming studies per se is outside the scope of this data report, it is important to know whether the key effect under investigation is present and consistent across all three samples (in this case the presence of congruency effects – i.e. are positive words evaluated faster when preceded by positive music than negative music, and similarly for negative words and music). To this end, there now follows a brief account of the crucial *Prime Valence*× *Target Valence* interaction. Prior to analysis, incorrect answers were deleted from the data set, and RTs were log transformed (Whelan, 2008). Timeouts (i.e., instances where no response was given or where the keyboard response was after 2000 ms had elapsed) were classed as errors and excluded from the data set. Next, the transformed data were subjected to 2 (Prime Valence) × 2 (Target Valence) linear mixed effects modelling. In all three samples, the interaction was significant, although the effect is much more clearly visible in the Lab and CS samples (Lab: $F(1, 62) = 10.92, p = 0.002, \eta_p^2 = .15$; Web: $F(1, 64) = 6.42, p = 0.02, \eta_p^2 = .09$; CS: $F(1, 66) = 10.04, p = 0.002, \eta_p^2 = .13$). The difference in effect sizes is evident also in Figure 6.1b – 1d.

6.3.5 Costs

A final consideration is of course cost. We have estimated the financial costs associated with the data collection based on 32 participants per sample. We carried out thirty-two experimental lab sessions totalling 8 hours. Paying participants £2.50 per session and costing a research assistant's time at £11.40 per hour (the lowest agreed rate of pay for graduate students at Anonymous for peer review University for the academic year 2018-2019) leads to an overall cost of 32*2.50 + £11.40*8 = £171.20. For the CS sample, we paid, £0.80 per participant, and data collection took roughly two hours. Including taxes, the additional fee to Prolific is 36%: 32*£0.75*1.36 + 2*£11.40 = £57.62. It is more difficult to estimate the cost of the Web sample. In total, the web version of the study was online for over two weeks, during which time it was necessary to occasionally repost the study on Reddit and check the number of responses. Furthermore, completing studies on SurveyTandem accounted for around 6 hours of researcher time (costed as 6*£11.40 = £68.40). There is also, however, an important trade-off to consider: although the cost of the data was significantly lower than for the Lab sample and comparable to the CS sample, the quality of the data is somewhat poorer in terms of its agreement with the Lab data, data wastage and visibility of the hypothesized effect.

6.4 Interpretation and usage

The aim of this Data Report is to provide support for the concept of online collection of reaction time data. Additionally, the authors are able to point out some limitations and benefits of the three types of data collection. Researchers might also find it useful to compare the three data sets using specific measures of importance in their research.

One of the most striking differences is the difference in attrition rates between the Lab and CS samples and the convenience web sample. Data from all of the participants in the Lab sample and from 94.5% of the CS sample was viable, whereas in the Web sample data from 84.6% of participants was considered viable. It is, however, difficult to know why this may be the case. One possibility is that web participants were less motivated in the absence of a concrete financial incentive. Another is that participants in the web sample may have felt less invested in the research as they were taking part remotely and had not met the researcher in person. Another option is that error rates were higher because participants were taking part in sub-optimal conditions, so there could have been environmental distractors that influenced the error rates. A final option is that one of the sites used for recruitment operates a system whereby researchers exchange participation in surveys; researchers acquire points by participating in other researchers studies. When they have accrued enough points, their study is in turn circulated to other researchers enrolled with the website. This comes with the risk that some researchers may have little intrinsic motivation to complete the tasks properly and allow the task to time out whilst still accruing points to allow for circulation of their own studies.

The results of the KS comparisons and visual inspection of the cumulative RT distributions suggest that, in principle, online collection of response time data can yield RT distributions that are comparable to those collected in a lab. However, much depends on the sample. In particular, the degree of alignment between the CS and Lab samples was much better than the alignment between the Web and Lab samples. It seems reasonable to assume that participants in the prolific sample were more motivated to complete the study than participants recruited via more traditional web methods, although it is not known whether this is a consequence of the fee paid to the CS sample or other factors such as curiosity or personality.

The per-target/prime pairing distributions from the Web sample differ significantly from the distributions recorded in the Lab. Given the significantly better agreement between the Lab and CS samples, it seems likely that this difference is a result of environmental or participant variables rather than browser or hardware differences.

Importantly, the hypothesized priming effect was present in all three data sets, with the caveat that the effect was much more visible in the Lab and CS samples as compared to the Web sample. Indeed, the Lab and CS samples resulted in almost identical η_p^2 effect sizes ($\eta_p^2 = 0.15$ and $\eta_p^2 = 0.13$ in the Lab and CS samples respectively, with a smaller effect size in the Web sample ($\eta_p^2 = 0.09$): it is noteworthy that it was still present despite the significant binomial test. One implication is that researchers who lack access to funding may still be able to collect usable data via a convenience web sample.

Overall, the data set presented here provides support for the use of web-based reaction time protocols. Researchers should, however, exercise care in their choice of participant pool. Where possible, researchers should opt for participant pools where they can be confident in the degree of motivation and engagement on the part of the participants. Moreover, researchers may wish to carry out confirmatory lab studies. The benefits of this approach extended also to faster data collection than was the case with the Web sample whilst being more cost effective than Lab data. The data presented here align with Hilbig's (2016) findings that, in principle, RT phenomena can be captured successfully online and that online RT methods can be used successfully in music cognition.

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics statement

This study was carried out in accordance with the recommendations of the Music Department Ethics Committee (University anonymous for peer-review) and approved by the same committee. In the case of the lab study, informed consent was given in writing; in the case of the online studies, informed consent was given via an online checkbox.

Author contributions

JA wrote the PsyToolkit code, oversaw the Lab and Web data collection, carried out the data analysis, and wrote the first draft of the manuscript. TE proposed the concept of comparing the three datasets, oversaw the CS data collection, and contributed to the authorship of the manuscript.

Acknowledgements

The authors would like to thank the PsyToolkit developer, Gijsbert Stoët, for his advice on using sound files in PsyToolkit.

Supplemental material

The supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02883/full#supplementary-material More precise details regarding the selection of prime stimuli and procedure for the

RT task are included as supplemental materials. Although strictly outside of the scope of this data report, we have reported the results of the congruency effects of interest in the priming task.

Data availability statement

The raw data, stimuli, PsyToolkit code and R scripts used to analyse the RT data can be found at https://osf.io/yhsqv/?view_only=87ae6378312c4a539fd6a5316c983afb

A copy of the priming task is available at https://www.psytoolkit.org/cgi-bin/psy2.5.4/survey?s=

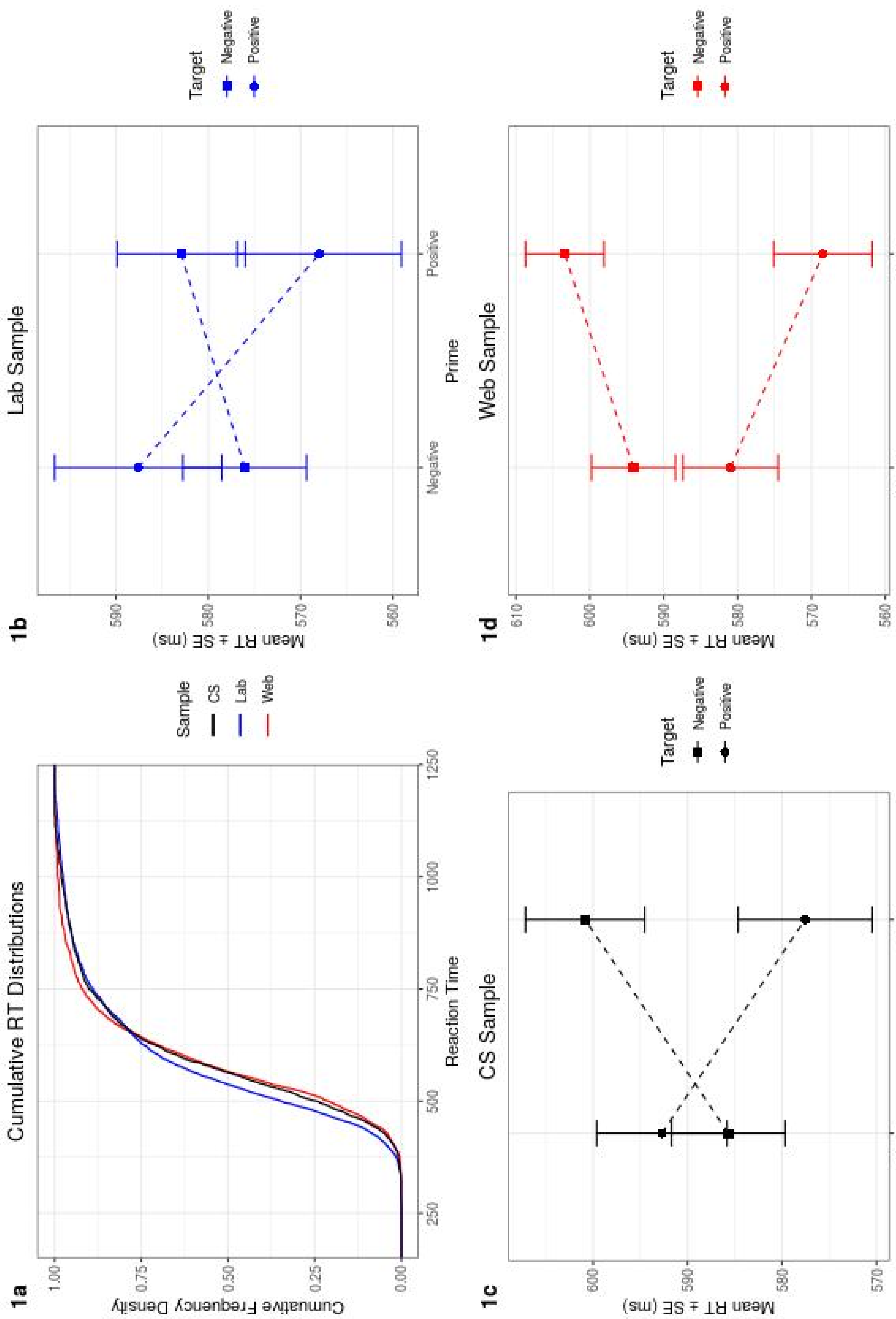


Figure 6.1: Cumulative RTs & Prime Valence×Target Valence Interactions for Lab, Web and CS Samples

Appendix

6.A Method for reaction time (priming) task

6.A.1 Materials & stimuli

32 musical clips, drawn from the list of excerpts used for musical mood induction compiled by Västfjäll (2001) and by Eerola and Vuoskoski (2011), were previously rated by a volunteer sample (n = 42, mean age = 37). The clips were rated on a scale of 1 to 7 for valence and arousal. The 16 clips (two per condition) that most clearly represented the four conditions Positive-High, Positive-Low, Negative-High, Negative-Low were chosen as stimuli for the main experiment. In common with Scherer and Larsen (2011), the auditory stimuli were each approximately 1000 ms in duration, where minor variation was allowed to maintain the musical integrity of the clips. Table S1 lists the musical excerpts, which are also available as an electronic appendix. The duration of 1000 ms is considered to provide the optimal

trade off in being sufficiently long to induce an affective response whilst avoiding decay in the priming effect (Bigand et al., 2005; Hermans et al., 2001). Musical excerpts rather than individual chords or short progressions have been chosen to mirror the ecologically valid approach taken by Scherer and Larsen (2011).

Excerpt	Valence	Arousal
The Good Life	Positive	High
Sousa - Stars and Stripes Forever	Positive	High
Holst - The Planets: Venus, the bringer of peace	Positive	Low
Shine	Positive	Low
Holst - The Planets: Mars, The Bringer of War	Negative	High
Lethal Weapon 3	Negative	High
Grieg - Aase's Todd	Negative	Low
Albinoni - Adagio	Negative	Low

Table 6.A.1: Music Primes by valence and arousal

The eight target words were taken from Warriner et al. (2013) and were chosen to represent the same distribution of valence-arousal pairings as the music clips. The target words are shown in Table S2. Together, these resulted in sixty-four target-prime (8 music prime × 8 target word) pairings. Target words were presented in the centre of the screen in white 40 point arial font on a black background.

Word	Valence	Arousal
Excite	Positive	High
Lover/Payday	Positive	High
Comfy	Positive	Low
Relax	Positive	Low
Arrest	Negative	High
Fatal	Negative	High
Dismal	Negative	Low
Morgue	Negative	Low

Table 6.A.2: Target words by valence and arousal

6.A.2 Procedure

Participants completed an affective priming task, in which target words were classified as positive or negative. Each target word was preceded by an audio clip. Items consisted of a fixation cross (450 ms), followed by the music prime. 450 ms into the prime, the target word was presented in the centre of the screen. Participants had a maximum of 2000 ms to classify the target word as positive (by pressing the m key) or negative (by pressing the z key). Participants completed two blocks: a ten-item practice block, and a sixty-four item experimental block. During the practice block, participant received a message after each item to inform them if their response was correct (correct!) or incorrect (incorrect response or too slow).

CHAPTER 7

Cross-modal transfer of valence or arousal from music to word targets in affective priming?[1]

Abstract

This registered report details a proposed study to consider how emotion induced in an auditory modality (music) can influence affective evaluations of visual stimuli (words). Specifically it seeks to determine which emotional dimension is transferred across modalities – valence or arousal – or whether the transferred dimension depends on the focus of attention (feature-specific attention allocation Spruyt et al., 2009). Two experiments were carried out. The first was an affective priming

[1]Published as Armitage, J., & Eerola, T. (2022). Cross-modal transfer of valence or arousal from music to word targets in affective priming? *Auditory Perception & Cognition*, *5*(3-4), 192–210. https://doi.org/10.1080/25742442.2022.2087451

paradigm that will allow for the orthogonal manipulation of valence and arousal in the both the words and music, alongside a manipulation to direct participants' attention to either the valence or the arousal dimension. Secondly, a lexical decision task allowed cross-modal transfer of valence and arousal to be probed without the focus of participants' attention being manipulated. Congruence effects were present in the affective priming task – valence was transferred in both the valence and arousal tasks, whereas arousal was transferred in the arousal task only. Contrary to predictions, the lexical decision task did not exhibit any congruence effects.

7.1 Introduction

The relationship between music and emotion is well established and has been the subject of much research activity in recent years (see e.g. Geethanjali et al., 2018, for a review). Music and emotion studies have considered a wide range of questions around emotion induction, personality traits, classification and emotion regulation. Importantly, musically-induced emotions do not exist solely within the context of music. Music also has the ability to influence emotional judgements about stimuli in other domains – a property that is utilised for instance in advertising and film music, or in compositional techniques such as word-painting or leitmotif. The present study seeks to probe more explicitly, via affective priming (Fazio et al., 1986), the nature of the transfer of emotions from music to other domains. In particular, it addresses the central question of what aspects of emotion can be transferred from music to influence judgements about stimuli in other dimensions.

7.1.1 Affective priming

The affective priming paradigm evaluates automatic responses to affective stimuli. Typically, two stimuli are presented in succession, and the second stimulus (the *target* is evaluated according to some affective criterion. The response to the target is thought to be influenced by the affective context provided by the first stimulus (the *prime*). For instance positive words may be classified more accurately (or quickly) when immediately preceded by a 'happy' sound than by a 'sad' sound; similarly, a negative word may be classified more accurately (or quickly) when preceded by a sad or angry sound than by a happy sound (see for example Scherer & Larsen, 2011). This phenomenon is known as a *congruency effect* or *priming effect*.

Priming studies have been used in music cognition for over thirty years, dating from Bharucha's series of seminal harmonic priming studies (for example Bharucha & Stoeckig, 1986, 1987). Although the present study does not address the mechanisms by which affective priming takes place, a brief overview of the two mechanisms that have been proposed as explanations of priming is necessary to address the question of which of valence or arousal is transferred. The first proposal (Fazio et al., 1986) is that priming effects are the consequence of spreading activation in the semantic network: exposure to a positive (negative) prime stimulus activates related concepts in the semantic network and consequently a succeeding positive (negative) target stimulus can be evaluated more quickly as the concept is already partially activated. An alternative explanation is that Stroop like interference inhibits responses when the prime and target stimuli are of opposing valence (the *incongruent* conditions) (see e.g. Klauer, 1997).

7.1.2 Transfer of basic emotion or valence-arousal

Broadly speaking, there are two commonly-used families of frameworks for studying emotion. These are the basic emotion models (for instance Ekman, 1992), and the dimensional models, such as the circumplex (i.e. valence-arousal) model put forward by Russell (1980). Ekman identified six different human emotions, i.e. families of affective states: *happiness, anger, fear, disgust, sadness* and *surprise.* Under this model, other emotions are considered to be combinations of these so-called basic emotions, so, for instance, jealousy can be considered a combination of anger, fear and sadness.[2] In the circumplex model, emotions are considered in the two-dimensional valence-arousal space. Broadly speaking, the valence dimension provides information about an emotion's pleasantness; the arousal dimension provides information about the degree of activation or tension. The circumplex model has been the most frequently applied framework to study emotional content of music (Eerola & Vuoskoski, 2012).

In addition to recent consideration of affect transfer described in terms of valence and/or arousal, there have been attempts to consider affective priming in terms of basic emotions (Boakes, 2010; Carroll & Young, 2005). Boakes compared the results for valence congruency with emotion congruency in three affective classification experiments and found a significantly greater facilitation effect for emotion congruence than valence congruence in two of the three studies (both using visual stimuli), with no significant difference in the third (lexical stimuli). However, most studies have used the valence-arousal framework, and it is possible to map basic emotions to the valence-arousal model. Furthermore, dimension-arousal Priming studies that have involved emotion more specifically have focused on valence rather than arousal as the transferred dimension of emotion (e.g. Bakker & Martin, 2015; Costa, 2013; Steinbeis & Koelsch, 2011). One exception is Marin et al. (2012) who argue that arousal rather than valence is transferred in cross-modal priming, whereas Scherer and Larsen (2011) and Liu et al. (2018) argue that both valence and arousal are transferred, but that valence makes a greater contribution to the priming effect than arousal. In general, however, there has been scant acknowledgement of the

[2]Ekman has since proposed an expanded list of basic emotions; other authors have proposed alternative lists, e.g. Izard et al. (1993)

role of arousal in the music priming literature. Furthermore, dimension-arousal is well-established as a model of music induced emotion (Eerola & Vuoskoski, 2011; Ilie & Thompson, 2006). Although Liu et al. (2018) use basic emotions as labels for stimuli, their careful manipulations of timbre were mapped to responses in valence and arousal. Therefore, the valence-arousal model is favoured in the present study.

7.1.3 Valence transfer

Most music priming studies (Hermans et al., 2001; Sollberger et al., 2003; Steinbeis & Koelsch, 2011; Yuan et al., 2014) have focused on transfer of valence, which mirrors the picture in priming studies more broadly (for instance Fazio et al., 1986). Scherer and Larsen (2011) argue that both valence and arousal are transferred in affective priming, but that the effect of valence is significantly greater than that of arousal. The main theoretical basis for valence transfer is provided by the spreading activation models (Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Hermans, De Houwer, & Eelen, 2001; Sollberger, Rebe, & Eckstein, 2003). In essence, information is classified simply as positive versus negative (or happy versus sad etc) and the corresponding memory nodes are activated; other concepts linked to these nodes are more accessible and consequently response times to congruent targets are quicker than to incongruent targets. Indeed, given the relatively large body of empirical work framed in terms of valence transfer, the congruency effect of valence does seem to be well-established. Furthermore, the theoretical models based around Stroop-like conflicts also seem to lend weight to valence transfer. However, as will be discussed below, considering transfer of emotion more fully potentially accounts for a greater proportion of the variance in priming effects than considering valence alone.

7.1.4 Arousal transfer

The main challenge, in music priming studies at least, to the notion of valence transfer is provided by Marin et al. (2012), who argue in favour of arousal transfer.

Outside of music studies, Zhang et al. 2012 and Hinojosa et al. (2009) argue in favour of arousal and the interaction of arousal and congruence, respectively. Marin et al. discuss arousal transfer in the context of excitation transfer theory (e.g., Zillmann, 1983). However, Zillman's theory suggests that arousal produced by an initial stimulus is slow to dissipate — there is consensus within the priming literature that the window for affective priming effects is very short, with effects decaying after about 250ms. Whilst Marin et al.'s protocol does use a short stimulus onset asynchrony, Zillman's theory seems to lend itself better to explaining phenomena such as aggression (Zillmann et al., 1972) and response to advertising (Mattes & Cantor, 1982) rather than affective priming, owing to the longer time scale that Zillman's effect works on. However, most authors (Scherer & Larsen, 2011; Sollberger et al., 2003) base their arguments on transfer of valence as discussed above. Interestingly, Scherer and Larsen do find a small main effect of arousal, but choose to frame their argument in favour of valence; it is not clear whether or not they considered or tested for an interaction. Storbeck and Clore (2008) review the available literature on the arousal dimension of affect as information, and discuss in detail the role off arousal in evaluation. Specifically, they argue that arousal tells us the urgency of the information.

A construct that has the potential to explain priming effects in terms of arousal is vigilance. Scherer and Larsen (2011) suggest this tentatively, but rely on more traditional explanations of priming in their discussion. Accepting vigilance as an explanation for priming would seem to suggest different responses to positive and negatively-valenced stimuli. However, if arousal is considered as the main vehicle for priming then the seemingly paradoxical facilitation effect in the positive-positive condition is resolved: high arousal is interpreted as threat in the first instance, irrespective of the valence, therefore the attentional spotlight narrows to look for other high arousal (interpreted as threatening) stimuli thus accounting the facilitation effects in both positive and negative congruent conditions. This explanation would be particularly suitable for the musical priming studies cited, given the relative simplicity of the auditory stimuli (usually major vs minor or consonant vs dissonant chords): positive chords convey excitement arousal and minor chords convey tension arousal (Lahdelma & Eerola, 2016). A model that gave primacy to the role of arousal would also potentially explain the inconsistency with research that suggests that negative affect inhibits cognitive performance.

7.1.5 The relationship between valence and arousal

Zhang et al. (2012) argue in favour of an interaction of valence and arousal in affective priming. Hinojosa et al. (2009) found no significant behavioural evidence in favour of a role for arousal in affective priming, but, when they considered ERP activity, they did find significant main effect of arousal congruency and an arousal congruency×target type interaction in the late positive component. It is feasible therefore that the interaction of valence and arousal is the key aspect in cross-modal affective priming, especially in light of Hinojosa et al.'s ERP results. Indeed, valence-arousal models of emotion could be consistent with the empirical basic emotion results discussed earlier. Curiously, Costa (2013) found that congruence effects were present when participants were asked to classify target words as positive or negative following valenced chords, but that prime valence did not lead to congruence effects using picture targets. He did however find significant facilitation in congruent register-affect conditions using faces as targets. He makes a tentative argument that arousal is transferred more effectively by higher register chords. Further support for the arousal transfer model is found outside of music cognition by Zhang et al., who similarly argue for a role for the interaction of valence and arousal in affective priming. However, in a purely lexical affective priming task, Hinojosa et al. (2009) found no significant main effect of arousal, but did find a significant arousal×congruence interaction. However, Storbeck and Clore (2008) review evidence which suggests that arousal narrows attentional focus, which would be consistent with Marin et al.'s account. Liu et al. (2018) matched prime and target stimuli for both valence and arousal. However, the focus of their research was eliciting affective priming effects via manipulations of acoustic properties rather than the dimensions of emotions. Nevertheless, they found significant priming effects in both behavioural and ERP results. Scherer and Larsen (2011) noted that their setup did not discriminate fully between valence and arousal transfer and suggested that further research be carried out to probe this issue more systematically. In particular, they suggested that, for their stimulus set at least, negative stimuli were notably higher in arousal than positive stimuli. Consequently, it was not possible to argue conclusively that one dimension had a greater effect than the other.

7.1.6 Feature-specific attention allocation

Rather than a simple transfer of valence or arousal, a strong alternative model of transfer is feature specific attention allocation (Spruyt et al., 2009, 2012). In simple terms, this model predicts that the transferred dimension will vary according to which dimension is attended to: if participants are making arousal evaluations, arousal will be transferred; if participants are making valence evaluations, then valence will be the transferred quantity. However, whilst this has been applied where the congruence categories under investigation have been affective versus not affective, it is as yet untested when the congruence categories have been two orthogonal dimensions of affect. Although Spruyt and colleagues present a strong case for feature specific attention, Becker, Klauer and Spruyt (2016), however, attempted, in an adversarial collaboration, to replicate Spruyt's findings – that for priming effects to be detected, attention must be drawn to the relevant dimension in the instructions – and failed to find significant priming effects, albeit using a pronunciation rather than an evaluative task. Becker et al. (2016) suggest that the original authors' results hinged on strong semantic relationships between the prime-target pairs. Therefore, as well has having significance in music emotion research, the present study will contribute also to the debate around the role of attention in affective priming.

Clearly, there is a strong consensus that both of the congruent conditions (positive prime - positive target, negative prime - negative target) facilitate reaction speed. Most studies have used relatively simple primes, e.g. major vs minor or consonant vs dissonant chords. *Prima facie*, this seems to support the idea of valence transfer. However, it still leaves scope to consider the question of valence vs arousal transfer if the nature of arousal is considered more closely. In particular, whilst Marin et al. (2012) considered arousal as one orthogonal dimension of affect, there is evidence to suggest that arousal can be further subdivided into excitement and tension arousal (Ilie & Thompson, 2006). Lahdelma and Eerola (2016) used this framework to consider emotion arousal in single chords and found empirical evidence to suggest that individual chords convey both valence and arousal, but that major chords convey excitement arousal, whereas minor and more dissonant chords are higher in tension arousal. Therefore, given the relative lack of consideration of the arousal dimension, many of the current studies do not intrinsically favour either an arousal or valence argument. Further studies are necessary to unravel the precise nature of the transfer.

There is strong agreement among authors that musical affective priming results in significant congruency effects: i.e., a positive stimulus facilitates categorisation of a positive target, and a negative stimulus facilitates categorisation of a negative stimulus. However, almost all music priming studies to date manipulate valence rather than arousal, whereas Marin et al. (2012) propose arousal as being the dominant dimension transferred in priming, albeit in a rating rather than a response time task. The present study uses the feature specific attention allocation (FSAA) framework proposed by Spruyt et al. (2009; 2012) to address this question.

The present study

The present study considers cross-modal transfer of both valence and arousal in two priming tasks. In Experiment 1, participants will complete a primed lexical decision task. In Experiment 2, participants will complete a word classification task, in which participants will categorise target words according to either their arousal or their valence properties. In both tasks, we will systematically vary both the valence and arousal properties of the prime and target stimuli. In this respect the design differs from existing studies, which have have for the most part used a binary division of auditory primes (positive/negative; major minor; resolved/unresolved). It will use musical primes that draw on a wider range of affects than has previously been the case. Despite the differences from previous studies, this experiment will incorporate aspects of replication of Scherer and Larsen (2011) using musical rather than non-musical primes. It will additionally consider more specifically correspondences in both arousal and valence with a view to evaluating the competing models proposed by Scherer and Larsen (2011), Marin et al. (2012), and Spruyt et al. (2009). With regard to feature specific attention allocation, existing studies have considered evaluation of affect versus evaluation of a non-affective category, where the affective classification has focused on valence only; this study will have separate valence and arousal classification tasks. To the authors' knowledge this is the first study to consider priming of valence and arousal as orthogonal dimensions exploiting feature specific attention allocation.

7.2 Experiment 1a and 1b: Valence and arousal priming

7.2.1 Participants

Brysbaert and Stevens (2018) suggest a requirement for 1600 readings per condition to achieve adequate power in mixed effect designs. Thus with 128 items (see procedure below) divided into four conditions, i.e. 32 items per condition, 50 participants were needed for the valence task, and 50 for the arousal task After the initial 50 participants completed the valence priming task and 50 completed the arousal priming task, i n light of deletion of incorrect item responses, an additional 8 and 7 participants were recruited in the tasks to ensure the 1600 item per condition threshold was met. Thus, the final sample sizes were 57 (valence task) and 58 (arousal task) Participants were recruited via Prolific, and were remunerated at an hourly rate of $11. All participants reported normal or corrected to normal vision and hearing, and were native speakers of English. Additionally, participants were right-handed (Hardie & Wright, 2014). Informed consent was given via an online check box, and the study was approved by the Ethics Committee of the host institution (Music Department, Durham University).

7.2.2 Materials

32 musical clips were selected from the list of excerpts used for musical mood induction compiled by Västfjäll (2001) and Eerola and Vuoskoski (2011). The clips were approximately 11 second in duration and represented the Western classical canon and film music genres. To establish which clips would be viable as primes for the present study, the clips were rated by a volunteer sample (n = 42, mean age = 37). The clips were rated on a scale of 1 to 7 for valence and arousal. The 16 clips (two per condition) that most clearly represented the four conditions, *Positive-*

Table 7.2.1: Music Primes for Affective Evaluation Task by Valence-Arousal (mean Valence and Arousal Ratings given in brackets*)

Positive-High	Positive-Low
Sousa – Stars and Stripes Forever (7.95, 8.02)	Shine (Tr 10) (5.88, 3.21)
Batman (Tr 18) (7.65, 7.62)	Pride & Prejudice (Tr 1) (6.18, 2.30)
Man of Galilee, CD 1 (Tr 2) (7.25, 7.25)	Holst – The planets: Venus (5.97, 3.64)
Tim Weisberg – The Good Life (7.71, 7.13)	Dances with Wolves (Tr 4) (6.42, 2.88)
Negative-High	**Negative-Low**
Holst – The Planets: Mars (3.68, 8.29)	Albinoni – Adagio (3.64, 3.64)
Cape Fear (1) (2.94, 6.98)	Grieg – Death of Aase (4.17, 3.06)
The Rainmaker (Tr 7) (3.24, 8.18)	Schubert – Quartet, D810., 2nd Mvt (3.15, 2.91)
Lethal Weapon 3 (Tr 8) (3.03, 7.93)	English Patient (18) (3.59, 3.37)

*Valence-arousal ratings were collected on a scale of 1 - 7 but here are normalised to 1 - 9 to correspond with ratings in Warriner et al. presented in Table 7.2.2.

High, *Positive-Low*, *Negative-High*, *Negative-Low* were chosen as stimuli for the main experiment and trimmed to roughly 1000 ms for use as primes. (see Table 7.2.1).

In common with Scherer and Larsen (2011), the auditory stimuli will each be roughly 1000 ms in duration (with minor deviation allowed to preserve the musical integrity of the clips) and will consist of musical extracts drawn from sixteen different pieces of music – four each to represent positive valence-low arousal, positive valence-high arousal, negative valence-low arousal, negative valence-high arousal. Table 7.2.1 lists the musical excerpts, which are also available at https://osf.io/qau7n/?view_only=52349d998411420ab0d3edc5084ef617. The duration of 1000 ms is considered to provide the optimal trade off in being sufficiently long to induce an affective response whilst avoiding decay in the priming effect (Bigand et al., 2005; Hermans et al., 2001). Musical excerpts rather than individual chords or short progressions have been chosen to mirror the ecologically valid approach taken by Scherer and Larsen (2011).

The sixteen target words (Table 7.2.2), two for each condition, were drawn from Warriner et al. (2013), and were matched for length. The words had previously been rated on a 9-point likert scale for valence, arousal and dominance by 1827 native English speakers.

Table 7.2.2: Target words for affective evaluative tasks grouped by valence-arousal (Valence-Arousal ratings given in brackets)

Positive-High	Positive-Low	Negative-High	Negative-Low
Climax (7.5, 6.8)	Gentle (7.4,3.2)	Rabid (3.0, 6.6)	Saggy (2.6, 3.0)
Lively (7.1,6.1)	Rest (7.9,2.2)	Hijack (1.8,6.1)	Flaccid (3.55,3.14)
Excite (7.8,6.6)	Comfy(7.2,2.9)	Arrest (2.3,6.7)	Dismal (2.6,3.3)
Snazzy (6.5,5.4)	Relax (7.8,2.4)	Fatal (2,6.8)	Morgue (1.8,3.5)

In each item in the priming task, the prime and the target stimulus either matched (congruent on both dimensions), partially match (congruent on one of arousal or valence) or mismatched (incongruent on both dimensions).

Alongside basic demographic measures, participants completed the one-item version of the Ollen Musical Sophistication Index (OMSI-1; Ollen, 2006; Zhang & Schubert, 2019). The affective priming task coded in PsyToolkit (Stoet, 2010, 2017) to allow for collection of robust reaction time data online via a local javascript (see Barnhoon, Haasnoot, Bocanegra & van Steenbergen, 2015, and de Leeuw & Motz, 2016, for discussion of online response time collection.) See Armitage and Eerola (2020) for a direct comparison of laboratory and online data collection of reaction time using music stimuli as well as the discussion of different samples.

7.2.3 Procedure

Following the procedure set out by Scherer and Larsen (2011), participants were initially informed that they were taking part in an experiment about the influence of music on reading. The whole experiment lasted around fifteen minutes. 50% of the participants completed the valence evaluation task; 50% completed the arousal evaluation task. Each task consists of a ten-item practice block followed by two experimental blocks of 64 items each, with a short pause in between. During the practice block, participants received feedback whether their response is correct or incorrect; no indication as to whether or not responses are correct was given during the experimental block. Within the blocks, the items were be presented in a random order.

Valence Evaluation Task

Participants were be presented with 1000 milliseconds of the auditory prime. During the auditory prime, the screen contained a fixation cross for 450 ms; after 450 ms, the target word was be displayed for 2000 ms, the window during which participants could respond. Participants were instructed to press 'z' if the word had negative associations and 'm' if the word had positive associations.

Arousal Evaluation Task

The procedure for the arousal evaluation task was be identical to that for the valence evaluation task, with the key difference that participants were instructed to press the 'z' key if the target word is low in arousal and the 'm' key if the target word is high in arousal. In addition, as the terms 'high arousal' and 'low arousal' may not be as intuitively clear to participants, participants saw an additional page of instructions explaining this in more detail prior to the experiment commencing: "Emotions can be classed as high arousal or low arousal. High arousal emotions are associated with energy, tension or activation. Examples of high arousal adjectives are awake, alert, restless. Low arousal emotions are associated with stillness or a lack of energy or activation, such as sleepy, tired, drowsy or serene. In this task you will classify words as high or low in arousal." (See Appendix for instructions given during the experimental task).

Statistical Analysis

All analyses were carried out in R (R Core Team, 2023) at $\alpha = .05$. Individual participants whose accuracy rate fell below 75% were excluded from the analysis. Timeouts (i.e. responses to individual items which exceed 2000 ms) were excluded. For RT analysis, incorrect answers were be excluded. Each participant's RT data will be fitted to an ExGaussian distribution (Ratcliff, 1993). In Individual RTs that are below 250 ms or greater than the 95th percentile of the ExGaussian distribution will be deleted.

RT distributions were analysed using a generalised linear mixed model (GLMM) fitted using a gamma function. The fixed effects were task (valence vs arousal) and congruence (congruent vs incongruent) and participants as a random factor, testing whether there is any difference in reaction times between congruent and incongruent conditions. Crucially, we considered the effect of interaction $Task \times Valence\ Congruence \times Arousal\ Congruence$ on reaction time (see hypotheses below) via a $2 \times 2 \times 2$ type III Anova. Planned contrasts with Bonferroni corrections for multiple comparisons were used to compare RTs in the different congruency conditions.

7.2.4 Experiment 2: Lexical Decision Task

The priming tasks outlined above tested which of valence or arousal primes responses to target words when participants' attention is focused on a specific dimension by the task instructions. However, the design of Experiment 1 did not allow us to determine the roles of valence and arousal in priming responses to target words in a more naturalistic context. Consequently, we include a second experiment in which the transfer of valence and/or arousal from primes to targets is considered without participants' attention being directed to either dimension.

Participants

Using Brysbaert and Stevens' (2018) figure of 1600 readings per condition, and 64 'real word' trials per participant split into four conditions, i.e. 16 trials per condition (see Design and Procedure, below), a minimum of 100 participants were needed to complete the study.

Materials and stimuli

The primes were identical to those used in Experiment 1. However, the target set included both 'real' target words and nonsense 'non-words'. The real words are identical to those used in Experiment 1. The non-words were generated using an online nonsense word generator and are four, five or six letters long, and consist of one or two syllables; the stimuli are available at https://osf.io/qau7n/?view_only=52349d998411420ab0d3edc5084ef617. The music primes are also identical to those used in Experiment 1.

Design and procedure

The outline procedure was similar to the valence and arousal classification tasks. The key decision, however, was whether a target forms a word or a non-word. Participants were presented with 1000 milliseconds of the auditory prime. During the auditory prime, the screen containrd a fixation cross for 450 ms; after 450 ms, the target word was displayed for 2000 ms, the window during which participants could respond. Participants were instructed to press 'z' in response to a non-word an 'm' in response to a real word.

Participants initially completed a ten-item practice block, during which they received feedback to indicate whether or not their response was correct after each item. This was followed by two experimental blocks of 64 items each, divided between 32 real words and 32 non-words. As with the valence and arousal tasks, the task was counterbalanced between participants to ensure all prime-target pairs are represented equally. During the experimental block no feedback was given as to whether or not the response to each item is correct.

Statistical analysis

Responses to non-words were deleted prior to analysis. The analysis broadly followed the strategy employed in Experiment 1. Again, the critical interaction under consideration was $Valence\ Congruence \times Arousal\ Congruence$. Planned contrasts analysis with Bonferroni corrections for multiple comparisons tested for differences in RTs in the following valence-arousal congruency conditions: *Congruent-Congruent, Congruent-Incongruent, Incongruent-Congruent, Incongruent-Congruent.*

7.3 Hypotheses/Expected results

All hypotheses apply to the affective priming task; only 1), 2) and 4) apply to the Lexical Decision Task.

1. Valence transfer: reaction times will be faster in the prime and target stimuli have congruent valences. Reaction time will be independent of any congruence in the arousal dimension.

2. Arousal transfer: reaction times will be faster in the prime and target stimuli are congruent in the arousal dimension. Reaction time will be independent of any congruence in the valence dimension.

3. Feature-Specific Attention Allocation hypothesis: in line with Spruyt et al. (2009, 2012), the interactions of $Task \times Valence\ Congruence$ and $Task \times Arousal\ Congruence$ will be significant and indicate that congruence effects will depend on task: responses to the valence task will exhibit congruency effects in the valence dimension but not the arousal dimension; responses to the arousal task will exhibit congruency effects in the arousal dimension but not the valence dimension.

4. Words that are congruent in both the valence and arousal dimensions will be classified as words more quickly than those which are congruent in one dimension only, which in turn will be classified more quickly than those which are not congruent in either dimension. In the case of Experiment 1, we predict more specifically that when there is congruence in one dimension only, RTs will be quicker when that dimension corresponds to the attended dimension in the task.

7.4 Results

7.4.1 Experiment 1

65 participants completed the arousal priming task and 58 completed the valence priming task. The item-wise deletion rate for the arousal task was higher than in the valence task, resulting in a slightly higher number of participants than expected in order to satisfy Brysbaert and Stevens' (2018) power criterion of 1600 readings per condition. We carried out the statistical analyses in line with the registered report plan. The critical three-way interaction $Task \times Valence\ Congruence \times Arousal\ Congruence$ proved significant, ($\chi^2(1) = 10.23, p = .0013$. Main effects and two-way interactions were assumed to be subsumed into the three-way interaction. The registered planned contrasts were carried out, revealing that in the valence priming task, there was a significant valence priming effect - i.e. RTs in incongruent conditions (mean RT = 560 ms, SD = 101 ms) were significantly slower than in congruent conditions (mean RT = 553 ms, SD = 102 ms), $z(\infty) = 3.700, p = .0001$, whereas no arousal priming effect was present, $z(\infty) = 1.24, p = .1105$: mean (SD) RTs in incongruent and congruent conditions were 558 (101) ms and 555 (102) ms respectively. In the arousal priming task, the predicted arousal priming effect was present, $z(\infty) = 8.903, p < .0001$, with RTs in the incongruent arousal condition (mean RT = 595 ms, SD = 133 ms) being slower on average than in the

congruent condition (mean RT = 581 ms, SD = 136 ms). In addition, the arousal priming task also revealed a significant valence priming effect - i.e. RTs for arousal classification were significantly shorter when the prime and target shared the same valence compared to when they did not share the same valence.

To test Hypothesis 4), we grouped RTs by task and by four levels of congruence: no congruence, congruence in the attended dimension, congruence in the unattended dimension, and congruence in both dimensions. We fitted a GLMM which was in turn submitted to an omnibus type III Anova. There was a significant effect of congruence. Planned contrasts revealed that there was no difference in reaction times between items where the congruence was in both dimensions and in the attended dimension only. However, RTs were significantly shorter in the conditions where the congruence was in the attended dimension compared to the unattended dimension. There was no significant difference in RTs between items where there was congruence in the unattended dimension compared to items where there was no congruence.

Exploratory analysis

To explore the presence of the valence priming in the arousal task further, we carried out an additional contrast analysis that was not included in the original registered report plan. We compared the valence congruency effect at the two levels of arousal congruency. Considering only those trials where the arousal levels of the prime and target were incongruent, there was no evidence of a valence congruency effect, $z(\infty) = 0.85, p = 0.3979$. However, when we considered the trials were the arousal levels of the prime and target were the same, there was a significant congruency effect - i.e. target words were evaluated more quickly when they shared the same valence as the music primes than when they did not, $z(\infty) = 3.00, p = .0027$.

Additionally (and with thanks to the anonymous reviewers for suggesting this analysis), we considered accuracy rate (AR). ARs were fitted to a GLMM, which was in turn subjected to a 2 (Task) x 2 (Arousal Congruence) x 2 (Valence Congruence) type III Anova. There was a significant two-way interaction of Task with Arousal Congruence, $Chi^2(1) = 7.95, p = .005$. Counter to expectations, post hoc t-testing revealed that there was a significant difference in AR in the valence classification task at different levels of arousal congruence: in the congruent condition, the mean AR was 89.2%, (SE = 0.58); in the incongruent condition, mean AR was 85.7%, (SE = 0.63), $t(115) = 4.62, p < .001$. All other two and three-way interactions were non-significant as were the main effects. The AR congruencey effect is shown in Figure 7.4.1

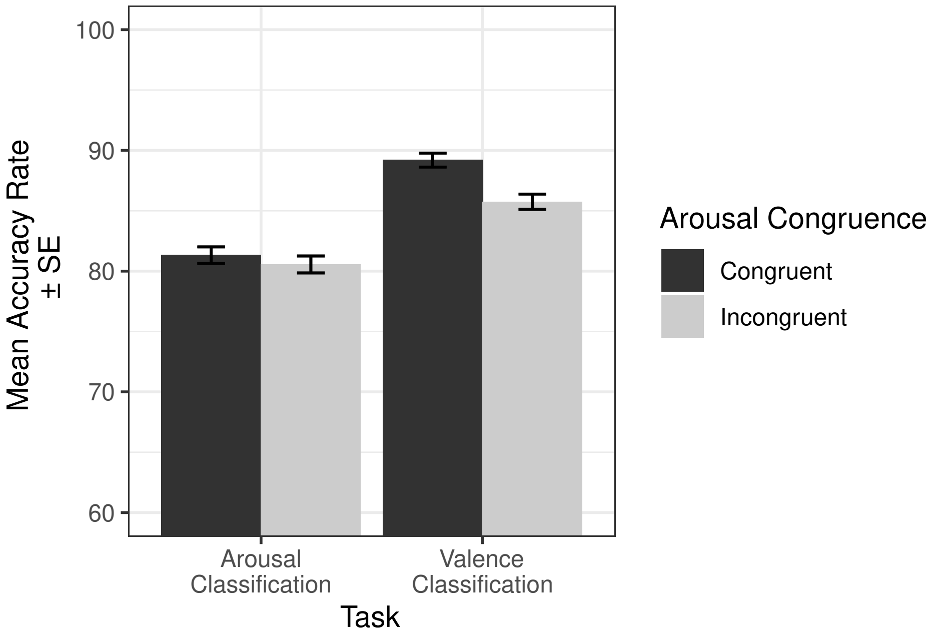


Figure 7.4.1: Accuracy rates for valence and arousal classification tasks by arousal congruence

7.4.2 Experiment 2: Lexical Decision Task

113 participants, who had not taken part in Experiment 1, completed the Lexical Decision Task; 5 failed to achieve the required accuracy rate, and so their data were removed prior to analysis. Following deletion of outliers (see Statistical Analysis, above), this left over 1720 readings per condition, meeting Brysbaert and Stevens' (2018) criterion for adequate power. The data were subject to a 2 ×2 type III

ANOVA. Contrary to the hypotheses above, no main effects of *Arousal Congruence* ($\chi^2(1) = 0.6196, p = .43$), *Valence Congruence* ($\chi^2(1) = 0.3705,\ p = .54$) were present; the interaction $Arousal\ Congruence \times Valence\ Congruence$ also proved non-significant, $\chi^2(1) = 0.047,\ p = 0.8382$.

Exploratory Analyses

Accuracy rates in the valence and arousal congruency conditions are shown in Table 7.4.1

Arousal Congruence	Valence Congruence	Accuracy Rate (SD)
Congruent	Congruent	88.5 (7.8)
Congruent	Incongruent	87.7 (7.9)
Incongruent	Congruent	88.0 (8.7)
Incongruent	Incongruent	88.0 (8.3)

Table 7.4.1: Mean (standard deviation) accuracy rates for valence and arousal congruency conditions

ARs were subjected to a 2 (Arousal Congruence) x 2 (Valence Congruence) Type III ANOVA. Both main effects and their interaction proved non-significant.

7.5 Discussion

Experiment 1 provided partial support for Hypotheses 1 and 3 – i.e. valence priming was evident in both the valence and arousal priming tasks. Arousal priming was present only in the arousal priming task. This perhaps suggests that the semantic networks have a automatic tendency to activate valence-related information, whereas – in this context – the arousal information is activated only when it is a requirement of the task. This asymmetric effect of valence and arousal congruency (i.e. valence congruency influence RTs on the arousal classification task but not vice versa) is perhaps analogous to (although in a very different domain from) the asymmetric effects of pitch on temporal position – variations of pitch influence judgements of temporal position whereas variations in temporal position do not influence judgements about pitch (Prince et al., 2009). However, the non-registered AR analysis suggests that the arousal information is activated but influences AR rather than RT. It is not clear why there should be a difference in how the arousal and valence information should influence responses. One possible explanation is that the valence classification task relates to a more fundamental conception of affect. This may not be strong enough to elicit AR priming effects independent of arousal, but when arousal is congruent, then the addition of valence congruence makes the alignment of the prime and target extremely close and hence sufficient to generate priming effects. Future research should utilise statistical methods that integrate accuracy rate and reaction time such as drift-diffusion modelling (Ratcliff, 1978). The non-registered contrast analysis perhaps points to the alternative explanation that, in the arousal task, increasing the degree of alignment in the affective information associated with the target in the prime causes an increase in the degree of facilitation. However, this effect is not apparent in the valence priming task.

There was partial support for Hypothesis 4: congruence in the attended dimension led to shorter RTs than congruence in the unattended dimension or no congruence. Interestingly, there was no RT advantage for instances where there was congruence in both dimensions compared to the attended dimension only. Again, this seems to lend weight to the argument that feature-specific attention allocation is an important factor in whether priming effects are present.

The absence of any of the hypothesised congruency effects in the Lexical Decision Task is surprising. To the authors' knowledge, this is the first time the difference between semantic and affective priming has been demonstrated in an auditory (and more specifically musical) context. We tentatively suggest that the absence of priming in the lexical decision task provides support for the Feature Specific Attention Allocation model (Spruyt et al., 2012) in that priming is only present in the tasks where the congruent dimension is referred to explicitly in the task. Alternatively, primed lexical decison tasks have been shown to be influenced by the associative as well as the semantic relatedness between the prime and target (e.g. Cańs, 1990; Perea & Rosa, 2002). It may be the case that there was an insufficiently clear associative relationship between the music primes and the target words to result in priming effects. However, it is possible that some aspect of the procedure could influence the results (e.g., the 450 ms SOA that has proved effective in the affective priming task could be sub-optimal in the lexical decision task; however, it should be noted that Holcomb and Anderson, 1993 found evidence of semantic priming at SOAs of 0 ms, 200 ms and 800 ms). Interestingly, this task was intended to give a sense of the influence of the affective content of music on word recognition without the influence of instructions that gave primacy to one dimension or another, i.e., an attempt to capture something more akin to a natural mode of listening. Therefore, it seems likely that a different paradigm may be necessary to truly capture the influence of emotional music on for instance the interpretation of lyrics.

The majority of evaluative priming research to date has focused on valence priming – it is possible that applying the same binary classification principle to arousal priming is less effective.

It should be noted that Hinojosa et al. (2009) argued that the interaction of arousal and valence plays an important role in affective priming.

Future research should consider whether the valence and arousal priming effects are unique to music primes or whether the effects present in this study are independent of modality. Additionally, the question of whether basic emotion provides a better framework for priming research should also be considered.

Data availability

The musical stimuli and PsyToolkit script are available at https://osf.io/hfbuk. Additionally, a draft of the valence priming task is available at https://www.psytoolkit.org/c/3.2.0/survey?s=CRsR9, the arousal task at https://www.ps ytoolkit.org/c/3.3.0/survey?s=Jngku and the lexical decision task at https://www.psytoolkit.org/c/3.3.0/survey?s=DazkM Raw .dat files, munged data in .csv format, R scripts and additional figures are available via the same repository.

Appendix

7.A Instructions for the priming tasks

The figures below give the instructions for the priming tasks.

Word Classification Task.

Instructions: In this task, you will see a series of words. If the word has negative connotations, press the '**z**' key, if the word has positive connotations, press the '**m**' key. Please respond within 2 seconds.

Each word will be preceded by a short audio clip.

You will now do a few practice attempts.

Press Space to start

(a) Valence Practice Block

Word Classification Task

As before, you will see a series of words. Press 'z' when you see a word with negative connotations and 'm' for positive connotations.

However, in this task there will be no indication whether or not your response is 'correct'

Press Space to start

(b) Valence Experimental Block

Word Classification Task

In this task you will see a sequence of words.

If the word is associated with **calm/still/relaxed** (ie low arousal) press the '**z**' key

If the word is associated with **energy/activation/excitement** (ie high arousal) press the '**m**' key

Each clip will be preceeded by a short audio clip: respond to the word not the sound.

You will now do a few practice attempts.

Press Space to continue

(c) Arousal Practice Instructions

Word Classification Task.

Again, you will see a series of words, preceded by audio clips. If the word is associated with **calm/relaxed/still**, press the '**z**' key, if the word is associated with **excitement/arousal/energy**, press the '**m**' key. Please respond within 2 seconds.

However, in this task there will be no indication whether or not your response is 'correct'

Press space to continue

(d) Arousal Experimental Instructions

Lexical Decision Task

You will now see a series of words on the screen. Some of the words will be real words. Some of the words will be nonsense words (pseudowords).

Each word will be a preceded by a short music clip.

If the word is a real word, press the 'm' key

If the word is a nonsense word press the 'z' key

Please answer as quickly and accurately as you can

You will now do a few practice attempts

Press *Space* to continue

(e) LDT Practice Block

Lexical Decision Task

As before, you will see a series of words on the screen. Some of the words will be real words. Some of the words will be nonsense words

Each word will be a preceded by a short music clip.

If the word is a real word, press the 'm' key

If the word is a nonsense word press the 'z' key

Please answer as quickly and accurately as you can. During this part of the experiment, you will not receive any feedback as to whether or not your answer is correct.

Press *Space* to continue

(f) LDT Experimental Block

Figure 7.A.1: Instructions for Priming Tasks

CHAPTER 8

Anxiety induced biases in auditory affective priming: A comparison of music and environmental sounds

Abstract

Trait anxiety is known to modulate responses to threat-related stimuli. We carried out two cross-modal affective priming tasks with auditory primes and word targets, comparing results for participants high in trait anxiety with those for participants low in trait anxiety. We made competing predictions for how trait anxiety would modulate the usual congruency effects based on Attentional Control Theory (Eysenck et al., 2007), Vigilance Avoidance (Williams et al., 1988), and the Information Processing model (Beck & Clark, 1988, 1997). Music primes elicited congruency effects - i.e., targets were evaluated more quickly when preceded by a prime that shared the same valence as the target. However, in this task we saw that trait anxiety did not influence responses in terms of either reaction time or accuracy rate. In the second task, which used affective environmental sounds, high

trait anxiety was associated with inhibited responses to positive targets following a negative prime, alongside globally slower reaction times compared to low trait anxiety. The pattern of responses was consistent with the impaired processing efficiency predicted by Attentional Control Theory. We discuss the results in the context of theoretical models of anxiety and suggest possible stimulus properties that account for the differences in results.

8.1 Introduction

Efficient processing of threat-related stimuli is critically important for human survival (e.g., Öhman, 2013). A large body of research (see, e.g. Bar-Haim et al., 2007) considers threat processing in healthy and anxious populations (e.g., Bar-Haim et al., 2007; Goodwin et al., 2017; Meyer et al., 2019), and there are several established theoretical accounts of the affective and attentional elements of threat processing (e.g., Beck & Clark, 1988; Bentz & Schiller, 2015). In particular, much research considers attentional processes and information processing in trait anxiety (TA) (Pacheco-Unguetti et al., 2010; Soyal et al., 2017; Stamps et al., 1979). Aubé et al. (2015) suggest that music activates the same threat processing mechanisms as fearful vocalisations. Research on strong emotional responses to music, such as chills or goosebumps, suggests fear or vigilance as a possible mechanism for these responses (Bannister, 2020; Huron, 2008). In clinical settings, music has also been shown to alleviate anxiety in certain situations (Bradt et al., 2013; Mallik & Russo, 2022), but the theoretical underpinnings of this phenomenon are not clear. More generally, there is little research on how music is processed in the context of threat perception and whether or not musical stimuli can activate anxiety-related processing biases. Here, we test whether brief exposure to positively or negatively valenced affective music and environmental sounds influences responses to an affective priming task and whether these responses are modulated by TA. Furthermore, we consider the degree to which the results are consistent with the TA models proposed by Beck and Clark, 1988, Williams et al., 1988, and Eysenck et al., 2007.

Despite its apparent simplicity, affective priming (Fazio et al., 1986) is a complex phenomenon that involves affective, semantic, and attentional networks. Typically, priming studies show evidence of congruency effects: targets are classified more quickly and/or accurately when they are preceded by a prime of the same valence as the target, but more slowly/less accurately when the prime and target are of opposite valences. However, some authors (e.g. Maier et al., 2003) report deviations from this pattern under conditions involving negative stimuli. According to the view taken by Hermans et al., 2003, it seems likely that there is an additional factor that influences the results of the priming tasks, which is present under negative conditions but not in positive conditions. One potential source of such variation in results is trait anxiety. TA is known to influence responses to negative stimuli (particularly threats). Additionally, the interaction of cognitive, attentional, and affective components of anxiety makes it an ideal framework for discussing affective priming in general.

As noted above, TA is associated with an attentional bias towards negative stimuli. Several accounts of anxiety provide a theoretical basis for the influence of anxiety on reaction time (RT) in priming tasks. Here, we will focus on models of anxiety which have the potential to explain the differing patterns of results outlined above. Beck and Clark's (1988; 1997) Information Processing Model of Anxiety, the Vigilance-Avoidance Hypothesis (Williams et al., 1988), and Eysencks (2007) Attentional Control Theory (ACT). We will also test whether or not music stimuli can activate anxiety-related processing biases reported by, e.g., Maier et al. (2003) and Hermans et al. (2003). The relationship between priming and anxiety is discussed in more detail below.

Beck and Clark, 1988, 1997 proposed that the initial perception of a stimulus (the orienting mode) is biased toward threat-related stimuli in anxious individuals. The second phase of processing a threat-related stimulus is the preparation mode, which involves activating the necessary schema to deal with the threat. During this stage of processing, Beck argues that all available information processing resources are deployed in processing threat-related information at the expense of other tasks. The final stage of processing a threat stimulus in Becks account is secondary elaboration, when the threat is appraised in terms of the individuals coping mechanisms and the availability of safety cues. Processing becomes more conscious with less dependence on automatic processes.

Williams et al., 1988 proposed the Vigilance-Avoidance Hypothesis. As with Becks information processing model, Williams et al. contend that TA is characterised by an early attentional bias towards threat stimuli. However, they argue that after initial perception of a threat stimulus, anxious individuals direct attention away from threat-related information in later stages of processing. They argue in favour of a mechanism that is only operational when competing stimuli are presented: vigilance-avoidance predicts that threat stimuli are not processed more quickly than neutral stimuli when the stimuli are presented in isolation, but rather that if threat and neutral stimuli are presented simultaneously, then the threat stimulus is attended to preferentially.

More recently, ACT (Eysenck et al., 2007) predicts that anxiety, both state and trait, is associated with greater vulnerability to distractor stimuli. In essence, the bottom-up attentional system is able to interfere with the top-down system. In the case of a priming task, the top-down task (target classification) would be subject to disruption by bottom-up signals created by the prime, particularly when the prime is threat-related. ACT suggests that disruption is particularly evident when working memory load is high, resulting in inhibited performance efficiency. When the demands for working memory are low, the efficiency of cognitive tasks performance is independent of the level of anxiety.

Previous research on priming and anxiety suggests a consensus around congruency effects when both the prime and the target stimulus are positive: targets are classified more quickly when preceded by a positive prime compared to when they are preceded by a negative prime. However, there is a less clear pattern of effects under congruent negative conditions (Dannlowski et al., 2006; Li et al., 2007; Maier et al., 2003). Whilst some authors found the expected congruency effect – negative targets are classified more quickly after negative primes – some have found reversed priming: negative targets are classified more slowly following negative primes. Li et al., 2007 carried out a priming study to assess the influence on TA on affective priming. As expected, they found standard priming effects. The size of the priming effect was found to correlate with TA, but crucially, this correlation was found only in the negative prime condition. Li et al. also considered EEG activity. The RT result was echoed in P1 activity, with the correlation only present in the negative prime condition. Maier et al. (2003) reported reverse priming effects in high TA participants exposed to high arousal targets. They argued that the (reverse) priming effects are a consequence of the interaction of TA and the activation level of the stimuli. They propose a dual spreading activation-inhibition mechanism: highly salient stimuli, rather than activating related concepts, actually inhibit them, resulting in reverse priming effects (see also Carr & Dagenbach, 1990). As an alternative, they point out that high TA participants may be motivated by fear of the consequences of low accuracy and respond slowly to ensure accuracy. In fact, the results are also consistent with the inhibited response efficiency to threat-related stimuli described in ACT. In contrast to this, Dannlowski et al., 2006 reported reverse priming in healthy controls, but priming effects in participants with anxiety disorder.

8.2 Experiment 1: priming with music

The present study uses a cross-modal affective priming task with two groups (low TA vs high TA). We predict that for the low TA group, we will observe standard congruency effects. Further, we predict that there will be some modulation of responses to negative stimuli in the high TA group. We test whether the results of the priming task are influenced by TA as predicted by the models of anxiety outlined above. As the nature of the predicted modulation by anxiety is unclear, the research seeks to assess three competing hypotheses:

1. As Rohr and Wentura (2022) consider that affective priming is a consequence of semantic activation in working memory, ACT predicts that processing efficiency will be disrupted following negative primes. This suggests that participants with high TA will display significantly slower reaction times when exposed to negative stimuli, i.e., ACT predicts a main effect of prime valence.

2. The Vigilance-Avoidance hypothesis predicts response times that are sensitive to the time course of the stimuli. Mogg et al. (2004) demonstrate that at short exposures (i.e., of 500 ms), participants high in TA had an attentional bias towards negative stimuli. This suggests that, at the SOA used here (450 ms) high TA participants will have an attentional bias towards negative targets after exposure to negative primes. This aligns with the reversed priming results reported by Berner and Maier, 2004. The hypothesis would be reflected in a three-way interaction between TA, prime valence, and target valence, and, in particular, shorter RTs in the negative-negative condition in the high-TA group.

3. Beck's information processing account predicts that trait anxiety will influence the accuracy rate; in particular, exposure to negative primes should lead to a situation being categorised as negative at the expense of positive aspects of the situation, i.e., accuracy rates in negative-positive conditions should be lower in the High TA group compared to the Low TA group.

8.3 Method

8.3.1 Participants

Brysbaert and Stevens, 2018 suggest that 1600 readings per condition are necessary for studies of this design to be sufficiently powered. As this required 25 in each of the Low TA and High TA groups, we aim for a target sample size of 60 participants to allow for attrition.

We first prescreened 700 potential participants for anxiety with the Beck Anxiety Inventory (BAI; Beck et al., 1988). Participants were recruited through Amazon MTurk and Prolific and remunerated at a rate of $0.65 (the screening task consisting of BAI took around 3 minutes). From the screening sample, the lowest and highest 25 quantiles of anxiety – excluding the upper extreme (5%) to avoid clinical or near-clinical cases – were recruited for the actual study that contains the priming tasks. This yielded two groups, low TA and high TA, each with a minimum of 20 participants. Participants received $3.65 for completing the priming task.

58 participants (25 male; mean age = 40 years, SD = 12.7) completed the priming task. The mean completion time was 20.4 minutes. Data from three participants were excluded as their accuracy rate fell below 75%, leaving 28 participants in the Low TA group and 27 participants in the High TA group.

All participants reported normal or corrected-to-normal vision and hearing, were native English speakers and were right handed (Hardie & Wright, 2014). Informed consent was given through an online check box and the study was approved by the Ethics Committee of the Department of Music, Durham University.

8.3.2 Materials

The musical stimuli were taken from Chapter 5 and were approximately 1000 ms in duration (1000-1045 ms, mean=1015 ms, where the minor variation was allowed to preserve musical integrity of the clips and avoiding cutting in the middle of onsets).

The stimulus set consisted of musical extracts drawn from sixteen different pieces of music – four each to represent positive valence-low arousal, positive valence-high arousal, negative valence-low arousal, negative valence-high arousal. Table 8.3.1 lists the musical excerpts. The duration of 1000 ms is considered to provide the optimal trade off in being sufficiently long to induce an affective response whilst avoiding decay in the priming effect (Bigand et al., 2005; Hermans et al., 2001).

Excerpt (Negative Valence)	Excerpt (Positive Valence)
Holst The Planets Mars	Shine
Schubert Death and the Maiden D.810	Pride and Prejudice
Grieg Death of Ase	Holst The Planets (Venus)
Albinoni Adagio	Dances with Wolves
Shine	Sousa Stars and Stripes Forever
Cape Fear	Batman
The Rainmaker	Man of Galilee
Lethal Weapon 3	Tim Weisberg The Good Life

Table 8.3.1: Music Primes

The sixteen target words (*Climax, Gentle, Rabid, Saggy, Lively, Rest, Hijack, Coma, Excite, Comfy, Arrest, Dismal, Snazzy, Relax, Fatal, Morgue*) were drawn from Warriner et al. (2013). The words were matched for length and arousal levels.

Trait Anxiety was measured using the Beck Anxiety Inventory (BAI; Beck et al., 1988). The BAI has demonstrated strong internal reliability ($\alpha = .92$) and strong test-retest reliability, $r(81) = .85$. The BAI is preferable to other anxiety instruments in having better construct validity as it is less subject to confounds such as depression.

The affective priming task was coded in PsyToolkit (Stoet, 2010, 2017) to allow for collection of robust RT data online. RT data collected online via crowdsourced samples has been shown to be of comparable quality to data collected in a lab (Armitage & Eerola, 2020; Kim et al., 2019).

8.3.3 Procedure

Participants initially completed the prescreening task (BAI). Following this task, selected participants were then contacted separately and invited to complete the next stage (the priming task). Participants completed the priming task approximately 24 - 48 hours after the prescreening task. Participants were initially informed that they were taking part in an experiment about the influence of music and personality on language processing. The experiment itself comprised a ten-item practice block followed by an experimental block of 256 items (16 primes $\times$ 16 targets). Participants received feedback during the practice block to indicate whether or not their responses were correct; no feedback was provided during the experimental block. Following the experiment, participants were presented with an online debriefing screen.

Participants were presented with 1000 milliseconds of the auditory prime. During the auditory prime, the screen contained a fixation cross for 450 ms; after 450 ms, the target word was displayed for 2000 ms, the window during which participants could respond. Participants were instructed to press 'z' on the computer keyboard if the word has negative associations and 'm' if the word has positive associations. Responses slower than 2000 ms were classed as timeouts. In each item, the prime and the target stimulus either matched (i.e prime and target had the same valence) or did not match (i.e., prime and target had different valences). Figure 8.3.1 summarises the procedure diagrammatically.

8.4 Results

8.4.1 Data analysis

Outliers were removed by participant by fitting an exponentially modified Gaussian distribution to the response time data and trimming any data in the 5% upper tails; response times less than 250 ms were also deleted. Timeouts and incorrect responses were removed prior to analysis. In total 11% of responses were removed from the

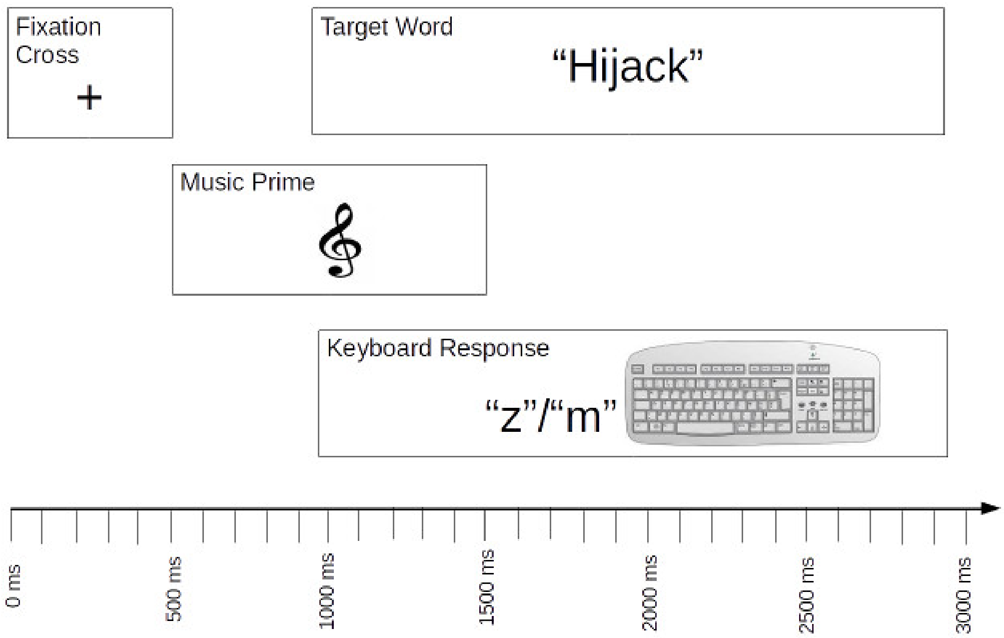


Figure 8.3.1: Procedure for auditory priming conditions

data set.

RTs were fitted to a Generalised Linear Mixed Model (GLMM) with a Gamma link function, with the fixed factors trait anxiety, prime valence and target valence and the random factor participant. The GLMM was then subject to a 2 (TA: High vs Low) × 2 (Prime Valence: Negative vs Positive) × 2 (Target Valence: Negative vs Positive) type III ANOVA. Where interactions of factors proved significant, main effects were assumed to be subsumed into the interaction. Where planned contrasts involved multiple comparisons, p-values were subject to Bonferroni correction. All statistical testing was carried out in R (R Core Team, 2023) at $\alpha = .05$.

8.4.2 Reaction time Analysis

Mean (SD) reaction times are reported in Table 8.4.1. ANOVA yielded a significant interaction of the factors Prime Valence and Target Valence, $\chi^2(1) = 18.42$, $p < .0001$. Planned contrasts revealed that negative targets were evaluated marginally faster following a negative prime compared to a positive prime, $z = 1.70,\ p = .08$ but the difference failed to reach significance; positive primes were evaluated significantly more quickly following a positive prime compared to a negative prime, $z = 4.40,\ p < .0001,\ 95\%\, CI = [5.37, \infty]$.

	Negative Prime Negative Target	Negative Prime Positive Target	Positive Prime Negative Target	Positive Prime Positive Target
High TA	604 (153) 90.0%	596 (149) 87.0%	608 (150) 88.7%	585 (149) 87.3%
Low TA	612 (121) 89.0%	602 (127) 88.8%	613 (110) 89.4%	595 (128) 90.3%

Table 8.4.1: Mean (SD) reaction times (ms) & accuracy rates for Experiment 1.

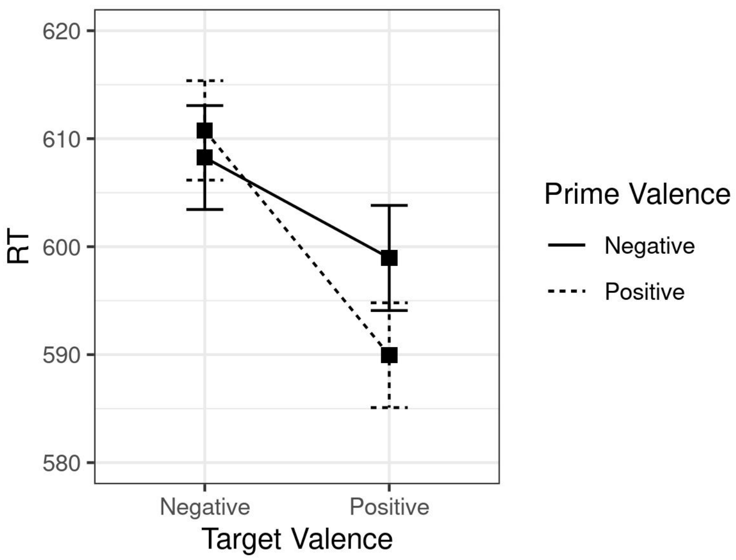


Figure 8.4.1: Interaction of Prime Valence and Target Valence

Contrary to the hypotheses above, the main effect of trait anxiety proved non-significant as did all two- and three-way interactions involving trait anxiety.

8.4.3 Accuracy rate analysis

Accuracy rates are reported in Table 8.4.1 and were subject to a 2 (Trait Anxiety: High vs Low) × 2 (Prime Valence) × 2 (Target Valence) ANOVA. All main effects and their interactions proved non-significant.

8.5 Discussion

As predicted, we saw a significant interaction of Prime Valence and Target Valence. As expected, we saw classical 'congruency effects' i.e., facilitation in the positive prime - positive target and negative prime - negative - target conditions versus inhibition in the negative prime - positive target and positive prime - negative target conditions, although the effect was weak in the presence of a negative prime. We did, however, see a main effect of target valence. This is potentially due to an overall positivity effect which saw positive targets processed more quickly than negative targets. The presence of congruency effects is consistent with the bulk of the music priming literature (Costa, 2013; Goerlich et al., 2011; Sollberger et al., 2003; Steinbeis & Koelsch, 2011; Tenderini et al., 2022).

We did not find evidence to support the predicted influence of trait anxiety on affective priming. Given that previous studies that have probed the link between trait anxiety and affective priming (Hermans et al., 2003; Maier et al., 2003) have found that there is an effect of trait anxiety on affective priming, it seems likely that the lack of an effect is a result of some property of the prime stimulus. Thus, further exploration is necessary to probe the nature of any anxiety-related biases in affective priming with auditory stimuli and empirically test which of the three models of anxiety (Beck et al., 1988; Eysenck et al., 2007; Williams et al., 1988) provides the best account of any such biases.

8.6 Experiment 2: priming with environmental sounds

Experiment 1 showed a clear facilitation-interference pattern in the sense that we saw faster reaction times in congruent condistions and slower reaction times in incongruent conditions. However, we did not see the expected influence on trait anxiety using affective music as primes. We carried out a further experiment to test the anxiety hypotheses a second time, this time using affective environmental sounds taken from the International Affective Digitised Sounds (IADS) database (Bradley & Lang, 2007). Affective environmental sounds were chosen as comparator stimuli because of their greater concreteness. Concrete (compared to abstract) threat stimuli are associated with increased physiological markers of anxiety in participants high in anxiety (Castaneda & Segerstrom, 2004). This allowed us to probe further the key question of whether the absence of an anxiety effect is something that is unique to music primes, or whether it extends to auditory primes more broadly.

8.6.1 Method

Participants

As in Experiment 1, participants were recruited from Prolific and were right-handed native speakers of English with normal or corrected-to-normal hearing and vision. 60 participants completed the priming task. 3 participants did not achieve the required accuracy rate and so were removed from the data set, resulting a sample size of 57 participants (26 female, 31 male; mean (SD) age = 41.7 years (11.8 years)). There were 30 participants in the Low TA group and 27 in the High TA group.

Materials and stimuli

Target words were identical to Experiment 1. The affective environmental sounds were taken from IADS (Bradley & Lang, 2007). The stimuli were taken from the same sound clips as Scherer and Larsen, 2011. The 11 - 14s clips were reduced to roughly 1000 ms for use in the priming experiment. The clips are listed in Table 8.6.1.

IADS Number (Positive Sound)	IADS Number (Negative Sound)
150 (Seagull)	255 (Vomiting)
151 (Robin)	276 (Female Scream)
224 (Kids)	279 (Attack 1)
351 (Applause)	284 (Attack 3)
353(Baseball)	285 (Attack 2)
802 (Native Song)	290 (Fight 1)
809 (Harp)	424 (Car Wreck)
816 (Guitar)	719 (Dentist Drill)

Table 8.6.1: IADS sounds used as primes in Experiment 3, adapted from Scherer and Larsen (2011)

Procedure

With the exception of substitution of the IADS stimuli in place of the music stimuli, the procedure for Experiment 2 was identical to that for Experiment 1.

8.6.2 Results

Data pre-treatment was identical to Experiment 1.

Mean (SD) reaction times and accuracy rates are presented in Table 8.6.2.

	Negative Prime Negative Target	Negative Prime Positive Target	Positive Prime Negative Target	Positive Prime Positive Target
High TA	640 (141) 90.6%	657 (182) 88.4%	656 (165) 89.8%	643 (173) 88.0%
Low TA	626 (141) 91.0%	622 (145) 89.5%	637 (143) 89.5%	618 (156) 90.3

Table 8.6.2: Mean (SD) reaction times (ms) & accuracy rates for Experiment 2

Overall, there was a main effect of trait anxiety, $\chi^2(1) = 31.2,\ p < .001$. Planned contrasts revealed that reaction times were, on average, slower in the High TA group (mean RT = 649 ms, SD = 172) compared to the Low TA group (mean RT = 626 ms, SD = 146), $z = 5.59,\ p < .0001,\ 95\%CI = [63.90, 159.69]$ As predicted, we saw a significant interaction of the factors Trait Anxiety, Prime Valence and Target Valence, $\chi^2(1) = 4.16,\ p = .041$. Planned contrasts showed that, for participants low in trait anxiety, when targets were preceded by a negative prime, there was no significant difference in RT between negative targets and positive targets. However, for participants high in trait anxiety, we saw that, when preceded by a negative prime, positive targets were associated with significantly slower reaction times than negative targets indicating significant interference effects, $z = 4.06,\ p = .0001,\ 95\%\,CI = [16.00, 39.92]$. Both the main effect of TA and the three-way interaction are represented graphically in Figure 8.6.1.

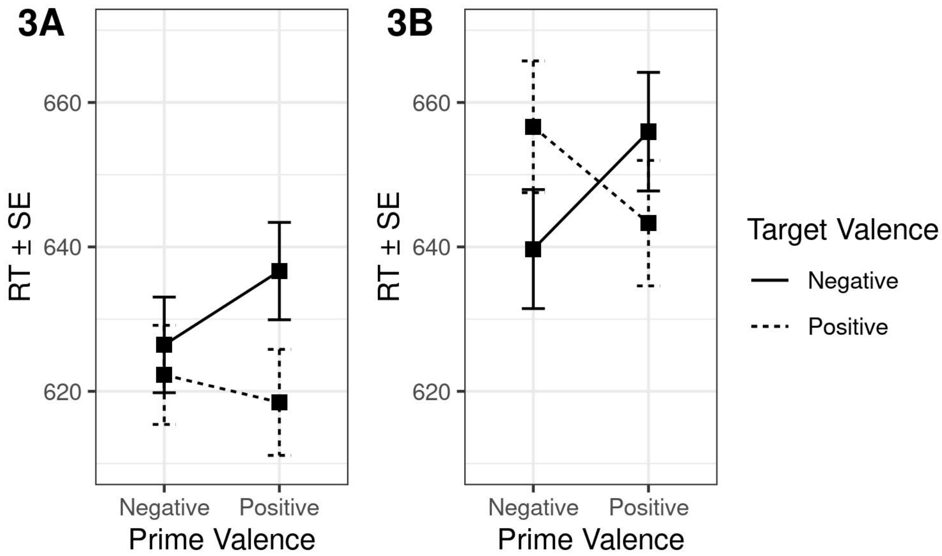


Figure 8.6.1: Interaction of Prime Valence and Target Valence for Low (3A) and High (3B) TA groups

8.6.3 Accuracy rate analysis

Mean (SD) reaction times are reported in Table 8.4.1 In common with Experiment 1, accuracy rates were subject to a 2 (Trait Anxiety: High vs Low) × 2 (Prime Valence) × 2 ANOVA. All main effects and their interactions proved non-significant.

8.6.4 Discussion

Experiment 2 considered how trait anxiety modulates affective priming in the case of affective environmental sounds. We saw that high trait anxiety was linked to overall slower reaction times. The slower reaction times for the High TA group are consistent with the consensus view that high trait anxiety is associated with slower reaction times on cognitive tasks than low trait anxiety (Pacheco-Unguetti et al., 2010; Soyal et al., 2017; Stamps et al., 1979). It is important to note that ACT predicts that *task efficiency* is diminished by high trait anxiety We saw that, in the high trait anxiety group, negative primes slowed down responses to positive primes (i.e., diminished the task efficiency), but that this was not the case in the low trait anxiety group. Moreover, the high working memory load associated with and affective priming task (Rohr & Wentura, 2022) meets the conditions for disrupted attentional control (Eysenck et al., 2007). Indeed, the absence of any significant result for accuracy rate is also consistent with Attentional Control Theory.

Considering next the, other two formulations of trait anxiety, we did not see an anxiety-related decrease in *task performance* – i.e. accuracy rate – as predicted Beck's Information Processing account of anxiety and priming, which predicts that targets will be erroneously categorised as negative. Additionally, the results do not support the Vigilance-Avoidance account, which predicts that, for the High TA group, responses in the negative-negative condition will be maximally disrupted (i.e., slowest). Overall, the inhibited reaction times in Experiment 2 seem consistent with an ACT account of anxiety.

Attentional Control Theory in essence suggests that bottom-up processes orient attentional resources to the negative prime stimulus. This limits the attentional resources that are deployed to the top-down word classification task and so processing efficiency, here indexed by reaction time, is inhibited. Critics of Attentional Control Theory point to the fact that ACT does not account for inhibition where the response should be highly automated, as is the case with an affective priming task.

Previous research into anxiety and priming has produced conflicting results – some authors have reported priming effects in negative conditions are amplified by trait anxiety, whereas other authors have found that priminig effects are reversed in high anxiety groups. Broadly speaking the present study is in accordance with Dannlowski et al. (2006) who found that priming was present in anxious participants but not healthy controls and Li et al. (2007) who found that there was a positive correlation between trait anxiety and the magnitude of the priming effect. However, the present result differs from Maier et al. (2003) and Hermans et al., 2003, who found that high TA was associated with reversed priming effects, i.e., for High TA groups, reaction times in the negative prime - negative target conditions were slowest. There are several possible reasons for the discrepancy between the results of the present study and those obtained by Maier et al. Firstly, the experimental paradigms are slightly different in that Maier et al. used a word pronunciation task rather than an evaluative classification task. Secondly, the difference in prime modality may have played a part. As both prime and target were represented visually (as written words) in Maier et al., it may be that a degree of unimodal interference is present in a word-word priming study that is not present in a cross modal study, e.g. a negative visual prime may trigger an aversive eye movement that interferes with the participant's ability to read the target word Onnis et al. (see, e.g., 2011, for discussion of time tracking of eye movements in response to negative stimuuli in anxious individuals) or slower disengagement from the negative prime (Okon-Singer, 2018). Similarly, Hermans et al. (2003) employed an image classification task with subliminally image primes. As with Maier et al. (2003), the reversed priming could be a consequence of eye movements. Additionally, subliminal presentation of the primes may bring about a different pattern of results from suprliminal presentation. Indeed, Hermans et al. posit that because concepts are activated only weakly in subliminal presentation, a kind of 'centre-surround' inhibition takes place, whereby concepts that are related but not identical to the prime are inhibited in semantic memory. Finally, and perhaps most crucially, Maier et al. specifically considered the activation level of the primes. In particular they found that the reversed priming effects were a consequence of activation level, occurring with the highest levels of activation of the affective representations. It is plausible that the activation level of the word primes employed by Maier et al. is greater than that generated by the auditory primes used in Experiment 2.

8.7 General discussion

The present study has shown that trait anxiety influences reaction time in auditory cross-modal affective priming, but that the influence is limited to environmental noises and not affective music. Whilst music has been used to induce state anxiety as part of other experimental paradigms and there is a wide-ranging literature on music performance anxiety, the present study is, to the authors' knowledge, the first to address the question of whether music can activate threat-related attentional biases or bring about processing inefficiency.

The different results for music and environmental sounds raises an important question: why does trait anxiety influence responses following one kind of auditory prime but not the other? The first explanation is activation level of the two stimulus types. It may be the case that the activation levels induced by the environmental sounds are great enough to induce anxiety-related processing inefficiency, whereas this may not be the case with the music stimuli.

An alternative explanation is that the difference in how music stimuli and environmental sounds is categorical, i.e, music stimuli may not carry threat-related information in the same way as the environmental sounds. I.e., whilst music is effective in eliciting a perceived emotion, it may be the case that it is not categorised as a threat stimulus at an early stage of processing. However, this is at odds with accounts of music-induced chills and goosebumps. For instance, several authors, such as Bannister (2020) and Huron (2008) contend that chills or goosebumps are a linked to threat-related information conveyed by music. One possible explanation is that these explanations are often linked to long-term musical structures and expectation violations, constructs that do not feature in the music primes, which had a length of approximately 1000 ms. Similarly, the result is to some extent in conflict with the findings of Aubé et al's (2015) that music activates the same processing mechanisms as fearful vocalisations.

A third explanation is that the concreteness of the affective environmental sounds is responsible for the presence of the priming effects with these stimuli but not music stimuli. It seems plausible that the sounds create a more realistic sense of threat, given that they are more directly related to events (such as violence, car accidents, or drill noises) that are unpleasant than the musical stimuli. Indeed, this is consistent with Castaneda and Segerstrom (2004), who found that threat stimuli high in concreteness are associated with greater vagal tension than more abstract threat stimuli.

Curiously, we saw no evidence of priming in accuracy rate results. This differs from the results in Chapter 7. The stimuli and procedure for the valence arousal task were identical to those in Chapter 7, so the result is not likely to be an artefact of differences in the experiments. One possible explanation is that RT is a more reliable index of priming than AR. Another possible explanation is that the more granular design (Prime Valence × Target Valence, rather than simple congruence) in this chapter masks an effect that is visible when analysed by congruence.

The present study uses English-speaking participants and uses Western tonal music as primes. It is therefore difficult to generalise the findings to other linguistic groups or to participants who are less familiar with Western musical idioms. In addition, the stimulus sets were chosen based on existing priming studies, (Armitage & Eerola, 2020; Scherer & Larsen, 2011). An explanation of the results based on activation levels could be tested more directly by manipulating the activation level of the primes.

Music is well established as a vehicle for emotion induction. However, in this instance, affective music stimuli have not tapped into anxiety-related processing biases as effectively as environmental sounds despite its prevalent use in anxiety reduction interventions (Bradt et al., 2013; Mallik & Russo, 2022; Nilsson, 2008). Future studies should consider how different types of auditory stimulus, spanning music and other areas, create induced and recognised emotions and whether different categories of stimuli provoke affective responses that are different in activation level or whether the responses themselves are categorically different. Furthermore, future research should address the relative lack of research on how trait anxiety mediates music-induced emotions.

Data availability

All data and analysis scripts are available at https://osf.io/mtzpw/?view_only=c0fa34ca804847b6bbee3bca2068128a

CHAPTER 9

Automatic responses to musical intervals: Contrasts in acoustic roughness predict affective priming in Western listeners[1]

Abstract

The aim of the present study is to determine which acoustic components of harmonic consonance and dissonance influence automatic responses in a simple cognitive task. In a series of affective priming experiments, eight pairs of musical intervals were used to measure the influence of acoustic roughness and harmonicity on response times in a word-classification task conducted online. Interval pairs that contrasted in roughness induced a greater degree of affective priming than pairs that did not contrast in terms of their roughness. Contrasts in harmonicity did not induce affective priming. A follow-up experiment used detuned intervals to create higher levels of roughness contrasts. However, the detuning did not lead to any further increase

[1]Published as Armitage, J., Lahdelma, I., & Eerola, T. (2021). Automatic responses to musical intervals: Contrasts in acoustic roughness predict affective priming in western listeners. *Journal of the Acoustical Society of America*, *150*(551). https://doi.org/https://doi.org/10.1121/10.0005623

in the size of the priming effect. More detailed analysis suggests that the presence of priming in intervals is binary: in the negative primes that create congruency effects the intervals' fundamentals and overtones coincide within the same equivalent rectangular bandwidth (i.e. the minor and major seconds). Intervals that fall outside this equivalent rectangular bandwidth do not elicit priming effects, regardless of their dissonance or negative affect. The results are discussed in the context of recent developments in consonance/dissonance research and vocal similarity.

9.1 Introduction

The contrast between consonance and dissonance is a vital feature of Western music. Consonance is typically perceived as *agreeable* and *stable* while dissonance, in turn, as *disagreeable* and *in need of resolution* (Tramo et al., 2001). Consonance/dissonance has both a vertical and a horizontal aspect: single isolated *intervals* (two concurrent pitches) and *chords* (three or more concurrent pitches) represent *vertical consonance/dissonance*, while the sequential relationships between these in melodies and chord progressions represent *horizontal consonance/dissonance* (Parncutt & Hair, 2011). The current research refers exclusively to the vertical aspect.

Empirical research concerning consonance and dissonance frequently relies on self-report methods, which come with a well-documented set of limitations (see e.g. Fazio & Olson, 2003, for review). Priming on the other hand captures participants' automatic responses to the stimuli, avoiding demand characteristics in a question that is underpinned by both physiology and culture (see e.g. Herring et al., 2013, for review). Previous studies using an affective priming paradigm have shown that valenced chords (e.g. consonant-positive, dissonant-negative) facilitate the evaluation of similarly valenced target words (see e.g. Steinbeis & Koelsch, 2011). Recent research (Lahdelma, Armitage, & Eerola, 2022) has found that this congruency effect is not present when intervals, as opposed to chords, are used as primes. Lahdelma et al. suggested tentatively that this finding was a consequence of the higher levels of roughness and/or harmonicity found in the four-note chords compared to the intervals, rather than as a consequence of the number of notes in the chords *per se*. Roughness is often seen as prevalent in the perception of dissonance but not in the perception of consonance see e.g. Hutchinson and Knopoff, 1978, as dissonant intervals contain less overall roughness than dissonant chords. Apart from roughness, another major acoustic factor related to consonance and dissonance is *harmonicity*, which has been demonstrated to contribute to perception of consonance in a variety of settings (McDermott et al., 2010). The study by Lahdelma, Armitage, and

Eerola, however, tested only four distinct intervals which poses a possible limitation.

The present study aims to disentangle the role of specific acoustic properties of the intervals, namely *roughness* and *harmonicity*. It considers automatic responses to these acoustic components as indexed in an affective priming paradigm (see *Affective priming*, below) rather than by self-report methods frequently employed in consonance-dissonance research.

9.1.1 Consonance and Dissonance

Acoustic Roughness

Historically, the acoustic and perceptual characteristics of consonance and dissonance were placed on a sound theoretical and empirical footing by von Helmholtz (1885) who proposed acoustic roughness as an explanation for why some musical intervals are considered dissonant and disagreeable. There is consensus that the sensation of roughness is caused by interference patterns between wave components of similar frequency that give rise to beating (see e.g. Hutchinson & Knopoff, 1978), which in turn creates the sound qualities which listeners typically perceive as unpleasant. Based on his findings of the relationship between roughness and dissonance, von Helmholtz derived that consonance in turn is the absence of roughness and concluded that roughness (or lack thereof) is the cause of both consonance and dissonance in music. Coming to a similar conclusion later, Terhardt (1984) proposed that the evaluation of consonance in isolated intervals and chords is mostly governed by *sensory consonance*, i.e. a lack of unpleasant features of a sound such as sharpness (the presence of spectral energy at high frequencies) and roughness.

The perception of roughness has a biological substrate, as beating occurs at the level of the basilar membrane in the inner ear when the frequency components are too close together to separate (see e.g. Tramo et al., 2001). This range is known as the *critical bandwidth* (Fletcher & Munson, 1933). According to Smith and Abel (1999, p. 21) "a critical band is 100 Hz wide for center frequencies below 500 Hz, and 20% of the center frequency above 500 Hz". A more recent formulation of this concept, the equivalent rectangular bandwidth (ERB, see Patterson, 1976) approximates the basilar membrane as being made up of rectangular band-pass filters. Fundamental frequencies processed within the same band-pass filter are perceived as acoustically rough. Compared to the critical bandwidth, the ERB typically encompasses a narrower range of frequencies (Smith and Abel, 1999). Although the ERB formulation is very much an approximation, it nevertheless provides a useful framework for considering aspects of vertical harmony.

The sensitivity to roughness seems to be present cross-culturally (see McDermott et al., 2016), but its appraisal differs significantly across musical styles and cultures: while a typical Western listener hears roughness as disagreeable, it is deliberately harnessed in the vocal practice of *beat diaphony* (known as *Schwebungsdiaphonie* in German literature) in for example the Baltic and Balkan regions of Europe (see e.g. Ambrazeviius, 2017) and in Papua New Guinea (Messner, 1981). Several models of roughness exist such as those by Hutchinson and Knopoff (1978), Vassilakis (2001), and Wang et al. (2013). These models, which are based on emulations of the human auditory system perceiving the sensation of roughness, largely agree on the amount of roughness in different intervals. Figure 9.1.1 shows how roughness varies for intervals over the octave from C_4 to C_5 as the distance between the two fundamentals increases by one cent – equal to 1/1200 of an octave.

While roughness was long accepted as a sufficient explanation for consonance and dissonance (see Helmholtz, 1885; Terhardt, 1984), later counterarguments were made for why it is not all-encompassing in explaining its underlying cause. First, perceptions of consonance and dissonance seem to remain when the tones of a chord are presented independently to the ears (i.e. dichotically), precluding physical interaction at the input stage and thus greatly reducing the perception of roughness (McDermott et al., 2010). However, it has been pointed out that beats could also occur centrally, within a binaural critical band rather than being based on cochlear interactions (Carcagno et al., 2019). Second, the perceived consonance of a chord does not seem to increase when roughness is artificially reduced by removing partials from complex tones (e.g. Nordmark & Fahlén, 1988). Third, it has been shown that participants with congenital amusia (i.e. a neurogenetic disorder characterised by an inability to recognise or reproduce musical tones) exhibit abnormal consonance perception but normal roughness perception (Cousineau et al., 2012). Recent research in roughness (Arnal et al., 2015, 2019) highlights the presence of roughness in, for example, alarm signals and suggests that acoustic roughness is responsible for activating salience-related aversive responses, particularly in the amygdala. Interestingly, Arnal et al. found that this activation was common to screams, alarm signals and – crucially for the present study – acoustically rough musical intervals. Additionally, Koelsch et al. (2018) found that stimuli high in acoustic roughness were associated with increased activation in the left planum polare and the orbital sulcus of the orbitofrontal cortex, a region associated with negative reinforcement (Kringelbach & Rolls, 2004).

Harmonicity

Another possible explanation that has been put forward to explain consonance and dissonance in addition to roughness is the acoustic property of *harmonicity* which denotes how closely the spectral frequencies of a sound correspond to a harmonic series. A single musical pitch is the combination of the fundamental frequency plus its overtones, which typically follow the harmonic series. If an interval is made of two tones which share a large proportion of their overtones (e.g. the perfect fifth) then the interval is high in harmonicity. Conversely, intervals whose overtones do not significantly overlap (e.g. the major seventh) are considered low in harmonicity. In other words, harmonicity posits that in consonant pitch combinations the component frequencies produce an aggregate spectrum that is typically harmonic i.e. it resembles the spectrum of a single sound. In contrast, dissonant intervals and

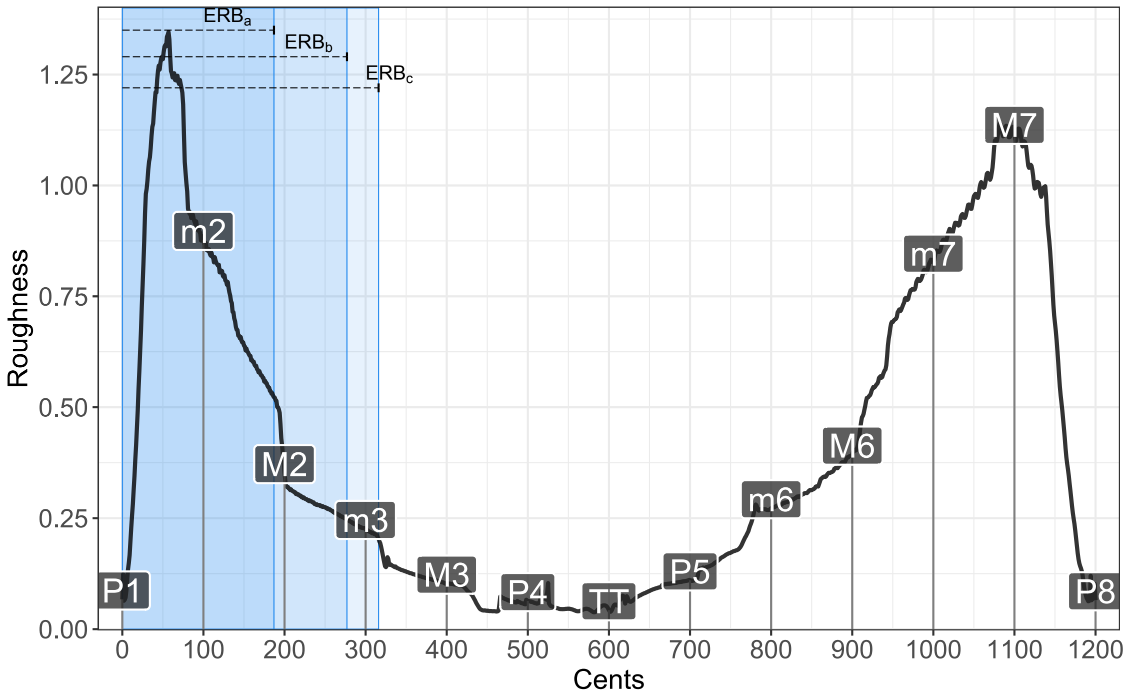


Figure 9.1.1: Roughness using the model by Wang et al., 2013 for intervals from unison to octave divided into 1200 cents. Frequency-dependent critical bandwidth boundaries are shown for the first three intervals (ERB$_a$ = m2, ERB$_b$ = M2, ERB$_c$ = m3) reflecting the different mean frequencies of the intervals in our design (see Stimuli) using the ERB bandwidths Moore and Glasberg, 1983.

chords produce an inharmonic spectrum (Cousineau et al., 2012).

Harmonicity offers a conception of consonance that is defined constructively rather than by the absence of roughness. Like roughness, harmonicity may have a biological underpinning, albeit at the central rather than peripheral level (Tramo et al., 2001), even if this is debated (Carcagno et al., 2019). Preference for high levels of harmonicity has been found to correlate with preference for consonance (McDermott et al., 2010), and it has been suggested that this link might be related to the advantages of recognising human vocalisation (Bowling & Purves, 2015). Another possibility is that, unless humans are born with a preference for harmonicity, it is acquired simply through exposure to natural sounds or to music (McDermott et al., 2010).

9.1.2 Affective Priming

Affective Priming is a common paradigm in both social and cognitive psychology, and is employed in diverse domains such as memory, social psychology, psycholinguistics and psychopathology as an indirect measure of attitudes (see e.g. Herring et al., 2013, for a review). In the affective priming paradigm, two stimuli are presented near-simultaneously. The extent to which the first (*prime*) stimulus influences responses to the second (*target*) stimulus in a classification task (for instance where the target is to be classified as positive or negative) is indexed by reaction time or accuracy rate. In particular, affective priming studies frequently consider *congruency effects*. Typically, in congruent conditions (i.e. when the prime and target stimuli share an affective feature such as valence) reaction times are faster than in incongruent conditions (i.e. when the prime and target stimuli do not share the affective feature.) Congruency effects have been associated with a range of auditory stimuli, for instance affective sounds (Scherer & Larsen, 2011) and music (Goerlich et al., 2011). Affective priming has been used in a small number of vertical harmony studies (Bakker & Martin, 2015; Costa, 2013; Lahdelma & Eerola, 2020; Sollberger et al., 2003; Steinbeis & Koelsch, 2011). There is a consensus amongst these authors that chords are associated with congruency effects, for instance dissonant (minor) chords paired with negative words and consonant (major) chords paired with positive words are associated with faster reaction times and/or higher accuracy rates than the converse. Despite the relative strength of agreement in the existence of congruency effects, there is relatively little discussion as to what drives these effects, either in terms of cognition or in terms of the specific acoustic properties of the chords. Steinbeis and Koelsch (2011) and Bakker and Martin (2015) suggest activation of affective concepts in the semantic memory and conflict-resolution in cross-modal integration respectively as possible causal factors on a cognitive level. The specific features of chords that may be responsible for congruency are not explored beyond modality, consonance-dissonance, register and numerosity. However, Steinbeis and Koelsch and Lahdelma, Armitage, and Eerola do tentatively suggest that acoustic roughness is responsible for presence of priming effects, the experimental manipulations varied consonance, modality and numerosity rather than varying roughness directly. Moreover, the authors' conclusions in favour of roughness cannot be considered conclusive owing to the absence of consideration of other factors, such as harmonicity.

9.1.3 The Present Study

The extent to which acoustical aspects (roughness and harmonicity) contribute to the perception of consonance/dissonance in intervals has remained contentious. The present study sought to explore the extent to which roughness and harmonicity influence automatic responses on a simple word evaluation task. As affective priming is an indirect measure, it is expected to yield valuable information on the importance of these individual factors' contributions to the perception of consonance and dissonance. The present study considered whether intervals can influence behaviour on a word evaluation task. An important question in cognitive psychology that has remained unclear in previous research is what property of the prime is responsible for activating the nodes in the semantic-affective network that lead to affective priming – perceived valence (i.e. positive or negative) or consonance/dissonance (teasing apart the components of roughness and harmonicity). The present study used ten interval pairs which were chosen to contrast in terms of their roughness or harmonicity (see Materials & Stimuli) to prise apart the differential contributions of these acoustic components of consonance and dissonance to affective priming. The intervals were chosen to maximise contrasts between the interval pairs in terms of roughness and harmonicity.

In a similar previous experiment (Lahdelma, Armitage, & Eerola, 2022), it was found that when positive or negative words were preceded by tetrachords (four concurrent pitches), there were significant congruency effects – i.e. positive words were classified more quickly when preceded by a consonant rather than a dissonant chord, and vice versa for negative words. Notably, these congruency effects were absent for intervals. Lahdelma, Armitage, and Eerola suggested that the context of exposure to tetrads might dampen responses to intervals and so the present experiments set out to examine whether diatonic intervals can drive priming effects without the confound of exposure to more complex chords. Ten separate within-subjects sub-experiments were conducted online, one for each pair of intervals (see Table 9.2.2). There were separate participants for each sub-experiment. Participants were asked to classify a sequence of emotional words as either positive or negative. Each word was preceded by the brief sounding of a musical interval. To evaluate the extent to which manipulating the amount of roughness and harmonicity can influence results of the behavioural task, each sub-experiment used two intervals, chosen for their contrast in roughness, harmonicity, or both. We predicted that intervals where there were high contrasts in harmonicity or roughness (or both) would be associated with congruency effects in reaction time (RT), whereas interval pairs which did not contrast greatly in roughness or harmonicity would not be associated with congruency effects.

Following the first ten sub-experiments, to probe the role of roughness further, we tested two interval pairs involving artificially detuned minor seconds (played with both the piano timbre and with the Shepard tone, see *Materials & Stimuli* for details). It is speculated that response to acoustic roughness confers an evolutionary advantage. For instance, alarm signals, whether in nature or man-made, are frequently high in roughness (Arnal et al., 2015). Indeed, specifically in the case of human vocalisation, roughness has been linked to perceived anger (Bänziger et al., 2015), screams (Schwartz et al., 2019), and infant cries (Koutseff et al., 2018). Consequently, it was predicted that these interval pairs which exhibited a higher difference in roughness than the standard diatonic intervals would elicit even greater priming effects than the high roughness contrast intervals.

9.2 Methods

9.2.1 Participants

Participants were recruited via Prolific Academic (https://prolific.ac), a crowd-sourcing platform targeted specifically for academic research. Following deletions (17 participants; see Data Analysis), 379 participants (197 female, 178 male, 4 other/prefer not to say, mean age = 36.2, SD = 12.4) completed the study. The target sample size was roughly 40 participants per sub-experiment in line with previous studies (Costa, 2013; Lahdelma, Armitage, & Eerola, 2022; Sollberger et al., 2003; Steinbeis & Koelsch, 2011; Tenderini et al., 2022). All participants reported corrected-to-normal or normal vision and were right-handed native speakers of English. 295 participants identified as non-musicians. Ethical approval was granted by the Music Department Ethics Committee, University of Durham. Informed consent was provided via an online checkbox.

9.2.2 Materials & Stimuli

Ten auditory stimuli were generated in total, eight diatonic intervals (equal tempered in common with Costa (2013) and Sollberger et al. (2003)) plus two intervals manipulated in tuning to maximise roughness. The diatonic interval pairs were chosen so as to maximise the contrasts in harmonicity and roughness; the diatonic intervals were classified as being high or low in contrast in roughness (ΔRoughness) and high or low in contrast in harmonicity (ΔHarmonicity), so there were altogether four conditions spanning the contrasts in roughness and harmonicity (see Table 9.2.2). The diatonic intervals were created in accordance with the procedure employed by Bowling et al. (2018) where the fundamental frequencies (f_0) of the pitches in each interval were adjusted so that the mean f_0 of both pitches in each interval pair was C_4 (261.63 Hz). Descriptors of roughness and harmonicity are given in Table 9.2.1. Roughness calculations were carried out using the model developed by Wang et al.

(2013). The analyses were duplicated using the Vassilakis (2001) and Hutchinson and Knopoff (1978) models, and with a composite model (mean roughness value of all three models). Harmonicity was calculated using the model by Harrison and Pearce (2018) which simulates the way listeners search the auditory spectrum for occurrences of harmonic spectra; harmonicity values under alternative harmonicity models operate similarly in this context and a composite model is detailed in the supplementary material. For the follow up sub-experiments, the two artificial intervals were combined with perfect fifths to create two additional interval pairs: a detuned minor second and a perfect fifth in Shepard Tones and piano timbre (hereafter abbreviated to s2/S5 and d2/P5 respecively). These interval pairs were classified as being 'Extreme' in roughness contrast.

As an additional diagnostic measure of dissonance, the first thirteen partials (fundamental plus twelve overtones) were extracted from each single tone from the intervals presented below. For each interval, we calculated the ERB about the mean of each pair of partials, using the formula derived by Moore and Glasberg (1983), generating 169 (i.e. 13×13) ERBs and checked whether the frequencies fell within this band. Table 9.2.1 details how many partial pairs for each interval fell within ERBs. The ERB was defined as the frequency band between the boundary suggested by Moore and Glasberg, 1983 and 10 Hz. Table 9.2.1 excludes those partial pairs that fell within 10 Hz because beating effects due to frequency differences of less than 10 Hz would be perceived as amplitude modulation or 'beats' rather than roughness see e.g. Roederer, 2008, p. 38; a further limitation of this measure is that very close alignments of overtones would contribute to a sense of harmonicity, for instance in the case of P5.

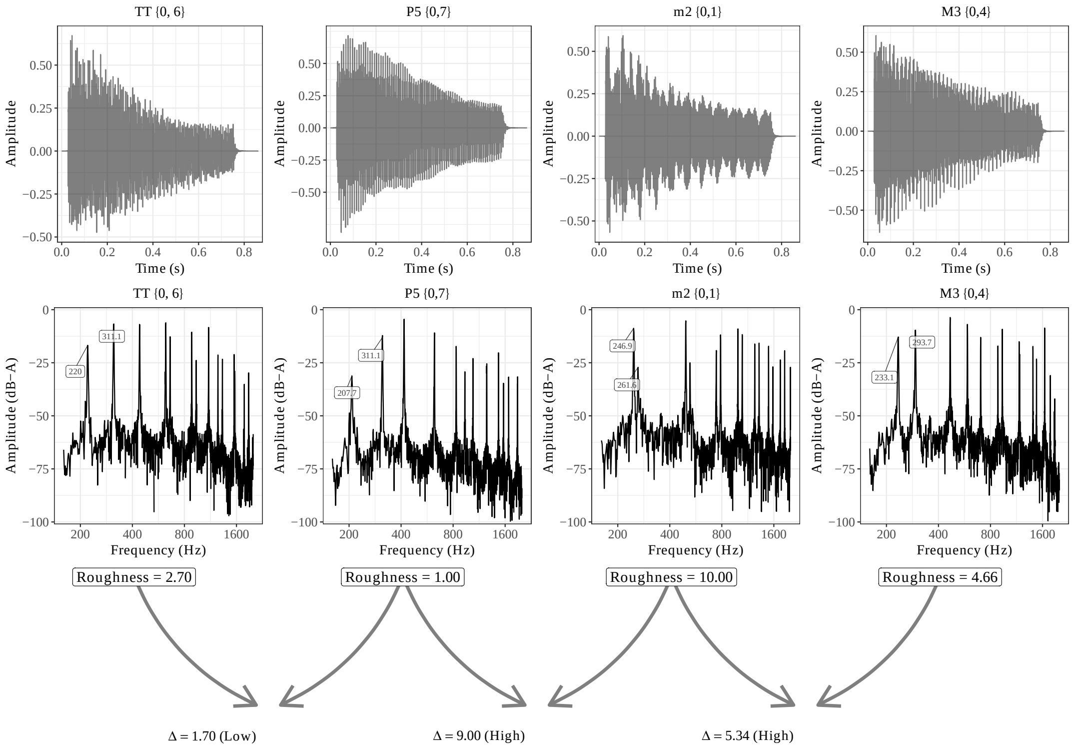
For the primes in the initial experiment, there were eight different pairings of intervals: m3/M3, m6/M6, TT/P5, m2/P5, M7/P5, m2/TT, M2/P5, and m2/M3. These interval pairings were generated with *Ableton Live 9* (a music sequencer software), using the *Synthogy Ivory Grand Pianos II* sample-based plug-in. For the

Table 9.2.1: Intervals, Notation, and Key Descriptors.

Interval (Abbr.)	Roughness	Harmonicity	Partials/ERB
Minor second (m2)	10.00	1.00	23
Major second (M2)	4.79	5.42	25
Minor third (m3)	3.66	2.79	18
Major third (M3)	4.66	4.72	15
Perfect fourth (P4)	3.64	10.00	14
Tritone (TT)	2.70	1.02	15
Perfect fifth (P5)	1.00	10.00	13
Minor sixth (m6)	4.75	4.72	14
Major sixth (M6)	5.98	2.79	15
Minor seventh (m7)	5.70	5.42	14
Major seventh (M7)	2.58	1.00	12
Minor second (detuned piano)	14.44	-	
Minor second (detuned Shepard)	10.11	-	
Perfect Fifth (Shephard)	-5.56	-	

piano interval pairs, the applied sound font was *Steinway D Concert Grand*. No reverb was used, and the intervals had a fixed velocity (65) in order to have a neutral and even sound. The artificial interval pairs comprised d2/P5 (detuned minor second) and s2/S5 (Shepard tones). In the d2/P5 interval pairing the minor second interval was created by taking a unison and detuning one pitch down by -90 cents; this procedure created a notably high amount of roughness when measured with the models by Vassilakis (2001) and Wang et al. (2013). The pairing s2/S5 was created using Shepard tones that have octave-spaced partials from 16Hz to 20kHz with cosine-curve-shaped spectral envelope (for the code used to generate the Shepard tones, see SI). The detuning of the minor second interval using the Shepard tone was created by shifting the odd partials above the fundamental upward and partials below downward by a detuning constant d. The constant was determined to yield a maximal roughness at $d = 0.024$ using the roughness models by Vassilakis and Wang et al. Table 9.2.2 shows the pairs of primes and the differences in their acoustic parameters. We used a median split to classify the differences in roughness and harmonicity as High or Low. The classification as High or Low ΔRoughness and High or Low ΔHarmonicity remained unchanged when we calculated composite

Figure 9.2.1: Amplitudes and Spectra for Sample Stimuli



Roughness and Harmonicity measures (see supplementary material).

The loudness of the stimuli was equalised by setting them to the same peak sone level. The sound files were converted to stereo (same signal in both channels) as 44.1 kHz, 32 bits per sample waveform audio files. The length of each interval was exactly 800 ms including a 10 ms fadeout. Figure 2 shows the spectra of a selection of stimuli alongside details of the calculation of the differences in roughness.

The target words were chosen from the database of affective norms for English words compiled by Warriner et al. (2013). All the words consisted of one or two syllables and were controlled for arousal so that there were four low-arousal and four high-arousal words in the positive and negative categories (Warriner et al., 2013). The negative words were *Flaccid (3.43), Hijack (1.84), Rabid (2.95), Coma (1.89), Saggy (2.62), Dismal (2.6), Arrest (2.33), Morgue (1.79)* (Warriner et al.'s valence ratings on a scale of 1 - 7 are given in brackets); the positive words were *Climax*

Table 9.2.2: Interval Pairs Tested for Priming Index (Difference in Roughness and Harmonicity given in brackets) and Number of participants.

Intervals	Δ Roughness	Δ Harmonicity	N
m3/M3	Low (1.00)	Low (1.97)	39
m6/M6	Low (1.23)	Low (1.93)	44
TT/P5	Low (1.70)	High (8.98)	37
M7/P5	Low (1.58)	High (9.00)	37
m2/P5	High (9.00)	High (9.00)	39
m2/TT	High (7.30)	Low (0.02)	38
M2/P5	High (3.79)	High (4.58)	33
m2/M3	High (5.34)	Low (3.72)	37
s2/S5	Extreme (12.61)	N/A	37
d2/P5	Extreme (15.67)	N/A	38

(7.53), Gentle (7.42), Lively (7.12), Rest (7.86), Excite (7.79), Payday (7.95), Relax (7.82), Comfy (7.25). The target words were presented in white size 40 Arial font on a black background. The priming task was coded using PsyToolkit (Stoet, 2017). Reaction time distributions collected via PsyToolkit and Prolific Academic have been found to be comparable to RT distributions collected in a controlled laboratory environment (Armitage & Eerola, 2020).[2]

9.2.3 Procedure

The experiment consisted of a standard word classification task with affective priming. Each item consisted of the prime (interval) presented simultaneously with a fixation cross for 250 ms. At 250 ms, the fixation cross was replaced with the target word. Participants were instructed to press the z key if the target word was negative and the m key if it was positive. The target word remained onscreen for 1500 ms; key presses greater than 2000 ms after the onset of the target word were classed as timeouts. Participants initially completed a 10-item familiarisation block, which was followed by the experimental block of 32 items. During the practice block, par-

[2]Owing to the Covid-19 pandemic we were unable to collect a confirmatory lab sample. However, we collected a small more tightly controlled web sample of 12 participants - all using MacOS, Chrome/Firefox browser and all of whom passed a headphone check prior to completing the experiment. The results mirrored those of the main experiment and are reported in the SI.

ticipants were informed whether or not their response was correct immediately after each item. No indication of accuracy was given during the experimental block.

9.3 Results

9.3.1 Statistical Analysis

Statistical analyses were carried out in *R* (R Core Team, 2023) with alpha = 5%. Each participant's RTs were fitted to a Gamma distribution using the R library *fitdistrplus* (Delignette-Muller & Dutang, 2015). RTs shorter than 250ms or slower than the 95th percentile of each participant's Gamma distribution were deleted (The mean 95th percentile was 936 ms, so a typical participant's RTs would lie in the range 200 ms to 936 ms). Similarly, incorrect answers and timeouts (i.e. RTs greater than 2000 ms) were also deleted prior to analysis. Data from participants who failed to reach a 75% accuracy rate was deleted from the analysis. Overall, 89.6% of data were retained. Data were analysed with a Generalized Linear Mixed Model (GLMM; model fitted using the glmer function from the R library *lme4*, (Bates et al., 2015)). GLMM analysis of RT data is discussed in Lo and Andrews (2015). Congruence, ΔHarmonicity and ΔRoughness were included as fixed factors; participants were treated as random factors. Standardised effect sizes are not reported for GLM models. The model used a Gamma family with an inverse link function. [3] To test the appropriateness of the Gamma distribution, each individual particpant's RT distribution was tested for goodness of fit against the Gamma distribution previously calculated (see above) via a Kolmogorov-Smirnov test. Of the 382 Kolmogorov-Smirnov tests, 11 returned significant results, $p < .05$ in each case. Finally, we carried out a binomial test ($r = 12$, $n = 379$, $p = .05$) which proved non-significant, $p = .98$, suggesting that significant Kolmogorov-Smirnov tests had occurred at

[3]The inverse link function was chosen as it offered the lowest AIC compared to logarithmic or identity functions

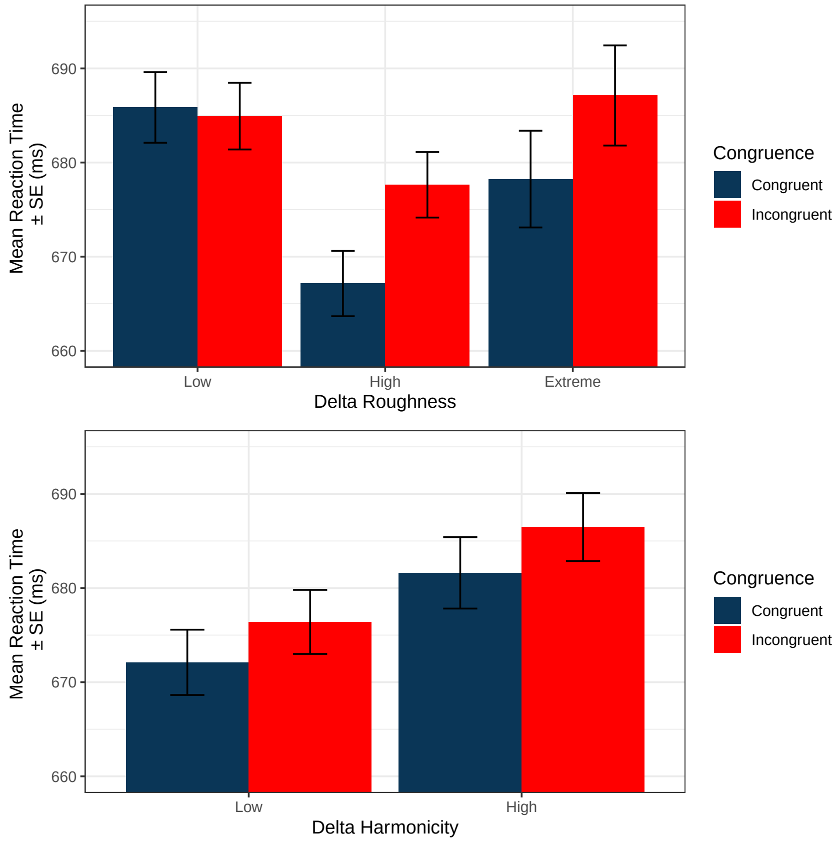


Figure 9.3.1: Mean Reaction Times for Low, High and Extreme levels of ΔRoughness and Low and High levels of ΔHarmonicity.

no more frequently than chance levels and that the gamma function provides an adequate representation of the distributions. The structure of the GLMM was $RT \sim \Delta Harmonicity \times Congruence + \Delta Roughness \times Congruence + 1|Participant$. The effects of this model were then subject to a type III ANOVA. The p-values in the subsequent planned contrast analysis were subject to Bonferroni corrections for multiple comparisons. Standardised effect sizes are not reported for the GLMM (Baguley, 2009).

9.3.2 Relationship Between Priming and Stimulus Features

Stimuli were grouped according to their acoustic properties: High vs Low difference in acoustic roughness (ΔRoughness), and High vs Low difference in harmonicity

(ΔHarmonicity) [4]. Response times in the High vs Low ΔRoughness difference conditions and the High vs Low ΔHarmonicity conditions were compared alongside the overall congruency effect (Congruent vs Incongruent conditions). As predicted, congruency effects were present in the High ΔRoughness condition. The key interaction ΔRoughness$\times$Congruence proved significant, $\chi^2(1) = 5.67, p = .02$. The main effect of$\Delta$Harmonicity was non-significant $\chi^2(1) = 0.54, p = .46$ as was the interaction ΔHarmonicity$\times$Congruence, $\chi^2(1) = 0.35, p = .55$. Owing to the presence of the interaction, we do not report on the main effects of ΔRoughness or Congruence. Planned contrasts were carried out for the four ΔHarmonicity$\times\Delta$Roughness conditions, with results as follows: The HighΔHarmonicity$\times$HighΔRoughness condition showed significant congruency effects (Incongruent: mean = 671 ms, SD = 170; Congruent: mean = 664 ms, SD = 166), $z(\infty) = 2.72, p = .01$. Congruency effects were also present in the LowΔHarmonicity$\times$HighΔRoughness condition (Incongruent: mean = 684 ms, SD = 155; Congruent: mean = 670 ms, SD = 151),$z(\infty) = 3.11, p = .004$. However, the High$\Delta$Harmonicity$\times$Low$\Delta$Roughness did not yield significant congruency effects, $z(\infty) = 0.492, p = 1.00$, (Incongruent: mean = 690 ms, SD = 184; Congruent: mean = 687 ms, SD = 169). Additionally, no congruency effects were present in the LowΔHarmonicity$\times$LowΔRoughness condition, (Incongruent: mean = 683 ms, SD = 169; Congruent: mean = 682 ms, SD = 175), $z(\infty) = 0.920, p = 0.920$. Mean response times per Congruency, ΔRoughness and ΔHarmonicity are shown in Figure 9.3.1. Within-subjects Wilcoxon tests on the congruency effects for individual prime pairs are reported in the supplementary material.

9.3.3 Roughness Manipulation

To probe the role of roughness further, we introduced two artificial intervals which were designed to test the influence of more extreme differences in roughness, d2 and

[4]Wang et al. (2013) and Harrison and Pearce (2018)

S2 (detuned minor seconds played with piano and Shepard tone timbres). Two further sub-experiments were carried out employing the same procedure but using these artificial stimuli as primes, with the expectation that the increase in ΔRoughness would, on average, increase the difference in response times between the congruent and incongruent conditions. As the key measure of congruency effects is the difference in response times, we carried out a simple linear correlation test on difference in roughness versus difference in response time between the Congruent and Incongruent conditions (referred to from hereon as *priming index* for brevity) for the expanded data set of ten interval pairs. The correlation was statistically significant, $r = .12, t(377) = 2.63, p = .02$. However, the low r value and visual inspection of Figure 9.3.1 suggested that, rather than a linear relationship, there is a step-like relationship – i.e. the increase in ΔRoughness is *not* associated with an additional increase in the priming index. To test this, the interval pairs were split into three categories by difference in roughness: Low (m3/M3, m6/M6, M7/P5, TT/P5), High (m2/P5, m2/TT, m2/M3, M2/P5), and Extreme (made up of the manipulated intervals d2/P5 and s2/S5). Owing to the clear a priori competing hypotheses (linear vs step-like dependence on ΔRoughness), we used a planned contrasts approach, where the dependent variable was the priming index. The planned comparisons were between the Low ΔRoughness and the combined High and Extreme ΔRoughness groups, and finally between the separate High and Extreme ΔRoughness groups.

For the first contrast, there was a statistically significant difference in the size of the priming index between the Low (mean index = -2.21ms, SD = 95.9 ms) and the combined High and Extreme groups (mean index = 22.4 ms, SD =83.5), $t(376) = 2.38, p = .04$. However, there was no significant difference in priming index between the High (mean = 10.9 ms, SD = 42.3) and Extreme (mean = 11.3 ms, SD = 40.9) ΔRoughness groups, $t(376) = 0.07, p = 1.00$, suggesting that increasing the level of roughness does not increase the strength of the automatic response, supporting the hypothesis that the relationship between ΔRoughness and priming

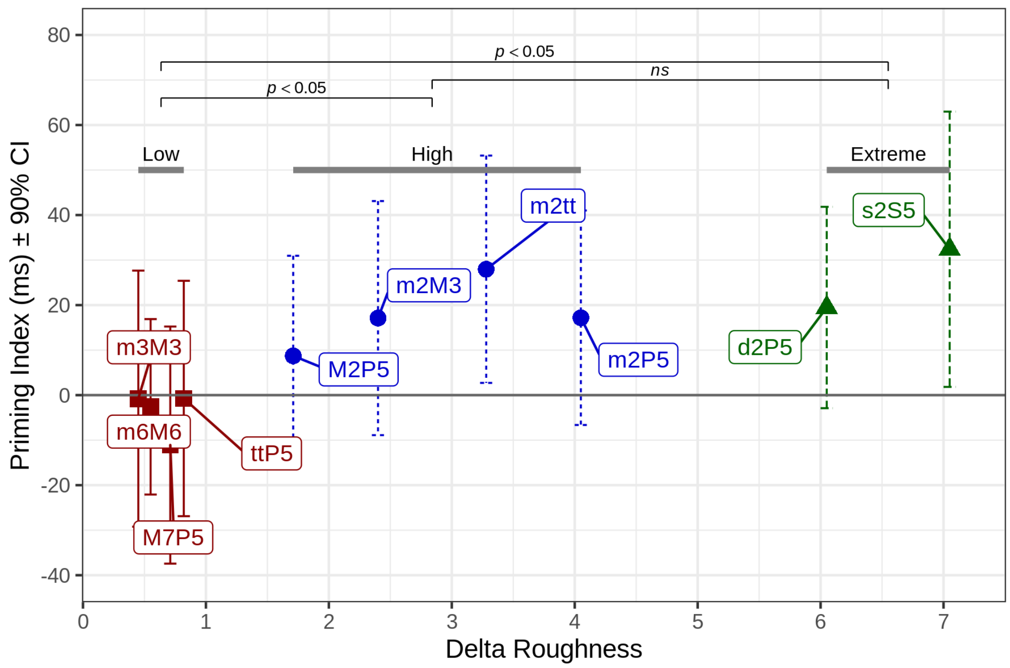


Figure 9.3.2: Difference in Roughness vs Priming Index for Diatonic Intervals. Three groups of pairings (Low, High, and Extreme Δ Roughness) highlighting the contrasts.

index is step-like rather than linear – that is, beyond a certain threshold, a further increase in the roughness contrast is not associated with a further increase in the priming index.

9.4 Discussion

We have demonstrated that interval pairs which differ by only a small amount in terms of acoustic roughness do not influence responses on a word classification task. On the other hand, intervals which differ more significantly in roughness do influence the responses. Notably, out of the two acoustic components that are seen as prevalent in the perception of consonance/dissonance, namely roughness and harmonicity (see Harrison & Pearce, 2020; Parncutt & Hair, 2011), roughness was the only component that influenced the responses. Harmonicity failed to reach significance, and consequently it seems likely that harmonicity contributes less to congruency effects in this setting than roughness - and the question of whether it contributes at all remains unresolved. Interestingly, the pairs m3/M3 and m6/M6 failed to influence responses to the classification task. This suggests that these interval pairs are not dissimilar enough in terms of roughness to elicit priming effects when presented by themselves, which is striking in the light that minor/major triads have been shown to be effective primes for negative/positive words in an affective priming setting (Steinbeis & Koelsch, 2011). Curiously, when used as the 'negative' prime in the pairing tt/P5, the tritone was not congruent with negative words, and was in fact an effective prime for positive words when paired with the maximally rough minor second interval due to its relatively low roughness (see Table 9.2.1). Given that the tritone carries conventionally a negative connotation in Western music (see e.g. Costa et al., 2000) and has historically been described as 'diabolic' (Partch, 1974), it seems important that further research should probe the relationship between cultural convention and acoustic components and how they could differentially influence automatic and appraisal judgements about consonance and dissonance.

Although several contemporary models of consonance and dissonance attempt to integrate the concepts of roughness, harmonicity and other non-acoustic factors (see e.g. Harrison & Pearce, 2020; Parncutt & Hair, 2011), the present study suggests that in this context of looking exclusively at musical intervals in an affective priming setting, roughness is the most important acoustic variable. Critically, the intervals that are associated with increased priming index, m2 and M2, fall within the critical bandwidth (see Patterson, 1976). The comparison of the minor and major thirds provides a particularly interesting case. The minor third falls theoretically just within the ERB. However, the priming index for the m3/M3 pairing did not differ significantly from zero. To probe this further, we considered whether the partials of the two notes in the interval co-occurred in the basilar membrane within the same ERB (see Table 9.2.1). This suggests that although the fundamentals lie within the same ERB for m3, the number of overtones lying within the same ERBs (18) is considerably less than is the case for m2 (23) and M2 (25), creating a much less rough effect overall. It should be noted that this approach does not account for relative amplitudes of the partials, and that the ERB formulation is an approximation – rather than a binary distinction, the degree of activation falls away more gradually. Moreover, it is not clear exactly how the summation of these interactions occurs. Nevertheless, it does provide a tentative explanation of why an automatic response was detected for other intervals in the ERB approximation (m2 and M2) but not for m3.

If we consider only the 'rough' intervals (i.e. the diatonic intervals m2, M2, and the artificial intervals d2 and s2), we did not detect a difference in the size of the congruency effects between interval pairs where the difference in roughness was high compared to extreme. This suggests that some qualitative difference exists between intervals that fall either within or outside this specific degree of ERB activation, but that beyond this threshold the degree of roughness does not influence the priming index. Indeed, the results of the present experiment are consistent with the suggestion by Scharf that "listeners react one way when the stimuli are wider than the critical band and another way when the stimuli are narrower" Scharf (1970, p. 196): intervals such as m2 and M2 create significant overlap in critical bands in both fundamentals and higher partials.

An important question is why it is the contrast in roughness, rather than harmonicity, that has a greater influence on the size of the priming index. Compared to harmonicity, roughness is uniquely situated in being associated with for example alarm signals and is thought to convey an advantage in enacting automatic responses (Arnal et al., 2015). One speculative explanation is that it is not dissonance *per se* that is pre-activating negative concepts. The human auditory system is well attuned to human vocalisations as they carry acoustic information about for example bodily states (see Pouw et al., 2020). Rough sounds are particularly salient and are associated with activation across large areas of the cortex (Arnal et al., 2019). It may also be that the salience of roughness, rather than its contribution to consonance or dissonance, is responsible for the priming effects. However, both in producing and attending to rough sounds, for instance cries of infants (Koutseff et al., 2018) or angry voices (Bänziger et al., 2015), sensitivity to roughness might confer an evolutionary advantage. This advantage explains why roughness in particular and not harmonicity is associated with an increase in the priming index, i.e. a stronger behavioural response to the valenced stimuli. It is plausible that the automatic response present in the priming index is driven by biological adaptation to acoustic factors: acoustically rough intervals activate responses associated with acoustically rough human vocalisations such as growls, screams, or cries. Indeed, recent research argues that responses to roughness in music have been exploited in the context of film music (Trevor et al., 2020). Notably, in the case of the major second interval this seems to happen quite subconsciously, if we go by the notion of composer/theorist Paul Hindemith who proposes that the major second sounds "almost consonant to our ears" (Hindemith, 1942, p. 85). On an empirical note, the major second interval has indeed been found to be perceived as more consonant than the minor seventh and major seventh intervals (Bowling et al., 2018).

Although the present study provides a behavioural method for tapping the roughness construct, it should be noted that at this stage the size of the effect is relatively small. Moreover, it is useful in indicating contrasts in roughness rather than as a direct measure of roughness. Nevertheless, roughness contrast is an objective method that taps into automatic perception as opposed to aesthetic judgements and it might mitigate semantic confounds which have been problematic for consonance/dissonance research in general as well as for specifically cross-cultural research endeavours into the question (see Bowling et al., 2017; Lahdelma & Eerola, 2020). Such an objective new method can help to investigate the appreciation of dissonance across musical cultures.

The present result provides, to some extent, an explanation of the observation that, in the absence of previous exposure to Western diatonic harmony, there is no preference for consonant intervals over dissonant intervals, although there is a small preference for large over small intervals (McDermott et al., 2016). The results of the present study suggest that the binary division is not whether an interval is categorised as consonant or dissonant, but on how much overlap there is in ERBs between the various partials - which is potentially higher in the case of for instance m2 or M2 than M7. The present study offers an explanation of automatic responses to acoustically rough intervals based on biological imperatives to respond to alarm signals present in human vocalisation. However, it should be noted that the emphasis in the present study is on *automatic responses*. Consonance and dissonance more broadly (e.g. cognitive or aesthetic appraisal) may well be underpinned by harmonicity or factors not considered in the present study such as cultural conventions. Future research should consider whether this biological imperative underpins dissonance perception more broadly or whether this sort of priming paradigm presents a special case. The role of culture in this effect also warrants further investigation, in particular whether it can be replicated with participants who have had frequent exposure to, for instance, *beat diaphony* in musical cultures that promote roughness for its aesthetic value (Ambrazeviius, 2017; Messner, 1981). The relationship between culture and harmonicity is difficult to disentangle. For the stimuli in the current chapter, the high degree of collinearity between harmonicity and cultural frequency meant it was not possible to control for cultural frequency in this instance (see e.g., Lahdelma, Eerola, & Armitage, 2022, for further discussion of the realtionship between harmonicity and familiarity). Additionally, future research could test the viability of exploiting acoustic roughness in alarm signals, for instance by using the major 2nd interval which provokes an automatic response in this priming context yet is not as unpleasant as other more extreme sounds.

Supporting Information and Open Data

Supporting information (different roughness and harmonicity models, Wilcoxon tests for congruency effects per interval pair, additional visualisations of spectra of stimuli) are publicly available, along with the reaction time data and analysis scripts at:]https://tuomaseerola.github.io/primingroughnessdata/. The PsyToolkit code and stimuli are available at]https://osf.io/zmjpd/

9.5 Acknowledgments

Partial financial support was provided by the Faculty Pro-Vice Chancellor's award at Durham University. Additional funds were provided by the Society for Education, Music and Psychology Research (SEMPRE) in the form of a Reg and Molly Buck Award awarded to the second author.

CHAPTER 10

Culture influences conscious appraisal of, but not automatic aversion to, acoustically rough musical intervals[1]

10.1 Abstract

There is debate whether the foundations of consonance and dissonance are rooted in culture or in psychoacoustics. In order to disentangle the contribution of culture and psychoacoustics, we considered automatic responses to the perfect fifth and the major second (flattened by 25 cents) intervals alongside conscious evaluations of the same intervals across two cultures and two levels of musical expertise. Four groups of participants completed the tasks: expert performers of Lithuanian Sutartinės, English speaking musicians in Western diatonic genres, Lithuanian non-musicians and English-speaking non-musicians. Sutartinės singers were chosen as this style of singing is an example of 'beat diaphony' where intervals of parts form predom-

[1]Published as Armitage, J., Lahdelma, I., Eerola, T., & Ambrazeviius, R. (2023). Culture influences conscious appraisal of, but not automatic aversion to, acoustically rough musical intervals. *PLoS ONE*, *18*. https://doi.org/https://doi.org/10.1371/journal.pone.0294645

inantly rough sonorities and audible beats. There was no difference in automatic responses to intervals, suggesting that an aversion to acoustically rough intervals is not governed by cultural familiarity but may have a physical basis in how the human auditory system works. However, conscious evaluations resulted in group differences with Sutartinės singers rating both the flattened major as more positive than did other groups. The results are discussed in the context of recent developments in consonance and dissonance research.

10.2 Introduction

10.2.1 Psychoacoustic vs cultural explanations of consonance and dissonance

The foundations of musical consonance and dissonance have received scholarly interest since the days of Pythagoras in ancient Greece (see e.g. Godwin, 1992), yet there is still no consensus as to whether they are a biological universal or culture specific. The notion of consonance and dissonance is fundamental to (Western) music theory, but there is still no fully accepted consensus of exactly what constitutes this categorical perception of musical sounds. Indeed, much of the empirical research into consonance and dissonance (hereafter referred to as C/D an implying exclusively simultaneous, not successive sounds) is confounded by related concepts such as pleasantness, valence and preference (for an overview, see Lahdelma & Eerola, 2020). Broadly speaking, explanations of C/D are typically rooted in some combination of psychoacoustics and culture. Psychoacoustic accounts of C/D explain it in terms of acoustic measures such as *roughness*, *harmonicity*, and *sharpness* (for overviews, see Eerola & Lahdelma, 2021; Harrison & Pearce, 2020), which are calculated based on the properties of the sounds themselves, and their interaction with the human auditory system. Perceptual roughness results when a complex tone displays a rapid amplitude modulation. This is a result of partials of a complex tone that are too close together to be resolved fully by the basilar membrane (see e.g., Hutchinson & Knopoff, 1978). Harmonicity in turn indicates how closely a sonoritys spectrum corresponds to a harmonic series (see e.g., Parncutt, 1989). Finally, sharpness denotes the energy at high frequencies which has also been identified as a predictor of C/D (see e.g., Eerola & Lahdelma, 2022; Fastl & Zwicker, 2007).

Sound combinations which are high in roughness are typically, though not universally, perceived as unpleasant on an aesthetic level. While Western listeners commonly perceive roughness as disagreeable (see e.g., Johnson-Laird et al., 2012; Lahdelma et al., 2021; Terhardt, 1974) it is harnessed for aesthetic ends (in moderate amounts) in the vocal practice of *beat diaphony* (also known in ethnomusicological literature by the German term *Schwebungsdiaphonie*) which refers to a two-part singing performance style where intervals of parts form predominantly rough sonorities and audible beats (Ambrazeviius, 2017; Ambrazeviius et al., 2015). It is present in diverse parts of the world, for example, the Baltic (Lithuanian Sutartinės) and the Balkan (e.g., Bosnian Ganga) regions of Europe (Ambrazeviius, 2017; Vassilakis & Kendall, 2008) as well as in Papua New Guinea (Messner, 1981), Nepal, Afghanistan, Ethiopia (see Vyinien, 2002), and the Indonesian islands (Kunst, 1942).

In addition to purely psychoacoustic accounts, there is empirical evidence to suggest that C/D is also influenced by culture; recent research has also pointed out that psychoacoustics and culture may overlap as in the case of harmonicity and cultural familiarity (see Friedman et al., 2021; Lahdelma, Eerola, & Armitage, 2022). In terms of cross-cultural experiments, Maher (1976) found that North Indian listeners rated dissonant intervals as markedly less tense than Western (Canadian) listeners, in line with the notion that musical intervals classed as highly dissonant in Western music are used more freely in North Indian classical music (Maher & Jairazbhoy, 1975). McDermott et al. (2016) in turn conducted fieldwork with the Bolivian Tsimané people, who are to a large extent insulated from Western cultural influences. The study concluded that the Tsimané are indifferent to C/D, although highly dissonant chords were omitted from the stimuli which somewhat hinders the results' generalisability; moreover, on closer inspection it is evident that the Tsimané did in fact have an aversion to the minor and major second intervals when presented diotically (simultaneous presentation to each ear). Despite this, McDermott and colleagues argued that on the basis of their results the foundations of C/D are purely cultural rather than psychoacoustic. More recently, Lahdelma et al. (2021) conducted fieldwork in remote Northwest Pakistan on how the minimally Westernised members of the Khalash and Kho tribes perceive harmony. Applying a wider range of C/D than McDermott et al. (2016) and using a direct selection task and pictorial representations to circumvent some of the semantic confounds that have been posing challenges for cross-cultural research, Lahdelma et al. found that there was an aversion to the highly rough chromatic cluster chord across both Western and the Khalash/Kho listeners, but that the preference for the consonance of the major triad was present only in Western listeners.

Taken together, there is an apparent conflict between the psychoacoustic predisposition for an automatic aversion to acoustically rough intervals as recently demonstrated by Armitage et al. (2021) with the preference for dissonant harmonic combinations that occur in the vocal practice of *beat diaphony* in many non-Western musical traditions (see e.g., Ambrazeviius, 2017; Vassilakis & Kendall, 2008). Armitage et al. (2021) argue that automatic responses to dissonant intervals are driven by interactions of fundamental frequencies in the basilar membrane – i.e., when the fundamental frequencies of a complex tone lie within the same critical bandwidth and the basilar membrane is unable to resolve them, resulting in an unpleasant sensation. The critical band has long held a central place in C/D theories, beginning with Helmholtz (1885), and underpins modern models of roughness. As demonstrated by Armitage et al. (2021), intervals high in acoustic roughness, such as the minor and major seconds, are processed automatically as negative in the case of Western listeners. Other intervals which have conventionally negative connotations in Western music, e.g. the 'sad' minor third, or the 'diabolic' tritone, are understood as negative as a result of culture. Such intervals are typically rated as unpleasant or negative in conscious self-report evaluations (see e.g., Bowling et al., 2018; Curtis & Bharucha, 2010) but not in the case of paradigms that probe automatic, unconscious responses to acoustic features (see Armitage et al., 2021). Interestingly, Maher (1976) questions the physiologically-based critical band theory's validity to account for the aversion to dissonant intervals based on the very finding that North Indian listeners perceived highly dissonant intervals as remarkably neutral compared to Canadians; indeed Maher (p. 271) proposes that "it does not seem likely that basilar membrane characteristics would be culture-specific".

It is crucial to note however that all of the previous cross-cultural experimental procedures into the question of C/D have used exclusively self-report measures – participants rated their impressions of the stimuli on a scale (see e.g., Maher, 1976; McDermott et al., 2016), or through preference choices and pictorial assessment (see Lahdelma et al., 2021). An important – and yet unresolved – question is whether these responses mirror or diverge from responses that happen more automatically without any opportunity for conscious appraisal of the stimuli. Automatic or implicit responses to consonant and dissonant stimuli have been investigated using reaction time (RT) and event-related potential (ERP) methods. Over the last 20 years, a number of studies have used affective priming techniques as a measure of attitudes to musical chords; priming studies using chord targets and word primes have found a strong consensus that hearing a dissonant three- or four-note chord facilitates processing of negative words, whereas hearing a consonant chord facilitates processing of positive words (Lahdelma, Armitage, & Eerola, 2022; Sollberger et al., 2003; Steinbeis & Koelsch, 2011). However, in the case of intervals the facilitation effects only seem to be present when the intervals used contrast maximally in roughness (Armitage et al., 2021). In particular, this effect has been observed when intervals that lie within the critical bandwidth (i.e., the minor and major seconds) have been used as the negative prime intervals, but not when other highly dissonant intervals such as the major seventh have been used.

Importantly, the questions of whether these automatic responses are subject to variation across cultures, and of whether and to what extent these variations mirror variations in self-report measures have yet to be addressed. Or to reiterate the sceptical view of Maher (1976) on the critical band theory, it is unclear whether mere exposure through cultural familiarity may override the physical and possibly universal response to the negative valence of dissonant intervals that lie within the critical bandwidth (see Armitage et al., 2021). One way of answering these questions is to utilise the priming paradigm with participants who are regularly exposed to what are considered dissonant intervals in Western music (and which are within the critical bandwidth), such as the minor and major seconds. Lithuanian Sutartinė singers is one such group that fulfils this criterion.

10.2.2 Sutartinės

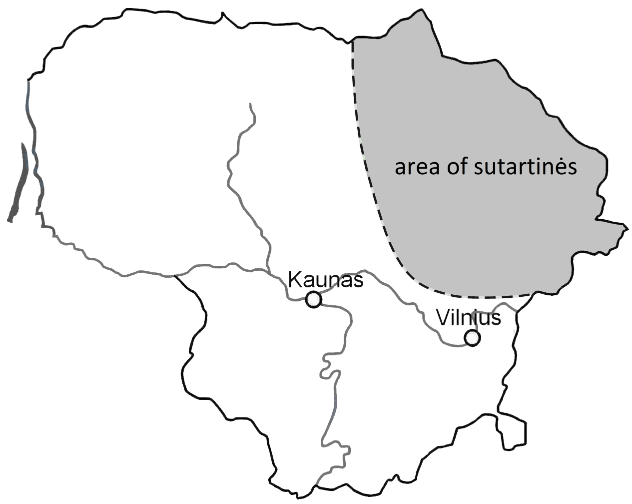


Figure 10.2.1: Map highlighting Sutartinės area within Lithuania.Republished from Ambrazeviius, 2021 under a CC BY license, with permission from Universität Bern, original copyright 2022.

The Lithuanian Sutartinė style is an unaccompanied vocal style, sung by female singers, originating in Northeastern Lithuania (see Figure 10.2.1). The name derives

Figure 10.2.2: Example of Sutartinė singing: An excerpt from "Myna, Myna, mynagaučio lylio" (adapted from Ambrazeviius et al., 2015, p. 221.)

from the verb *sutarti* which means to agree, to be in concord (Ambrazeviius, 2017). It is characterised by a narrow melodic range, the prevalence of dissonances (particularly seconds, see e.g. Ambrazeviius and Winiewska, 2009) and heterophonic textures in three parts (although two- and four-part Sutartinės are also common). Figure 10.2.2. provides an example of a typical vocal texture in this style. The major and minor seconds are utilised for expressive purposes; notably, the rough 'clashes' of the seconds are positively connoted and are appreciated for their 'bell-like' sound (Ambrazeviius, 2017) and this connotation is also known in the vocal traditions of Bulgaria, Serbia, Bosnia-Herzegovina, Macedonia, South Albania, Romania, and Northern Greece (Vyinien, 2002). This analogy presumably arises from similar psychoacoustic qualities including close partials, beats, attacks, and frequency range (see Ambrazeviius, 2017; Brandl, 1989). Although vocal Sutartinės are more common, there also exists and instrumental Sutartinės tradition (Nakien, 2003) that employs traditional Lithuanian woodwind instruments.

Sutartinės reached near-extinction in the mid 20th century (Ambrazeviius & Winiewska, 2009; Rainait-Vyinien, 2012), but have since gained popularity amongst folk music groups as a part of the broader revival in interest in the folklore and music of the Baltic countries Strmiska, 2005. Ambrazeviius and Winiewska (2009) probed the tonal hierarchies of a corpus of Sutartinės and reported that Sutartinės singers exhibited different cognitive representations of the tonal hierarchies, for instance attributing greater salience to the two central notes in the Sutartinės scale – akin to a "double tonic" effect – compared to controls and they also attributed much lower salience to the more distant notes form the tonic compared to controls. This is consistent with Krumhansl and Castellano (1983), who provide a schema-based account of music perception. In particular, they present the idea that there is a relationship between perception of musical events and a listeners knowledge of the underlying structures of the musical style. Specifically with regard to Sutartinės, Ambrazeviius and Winiewska's findings suggest that listeners with expertise in Sutartinės have a more distinct cognitive representation of the tonal hierarchies of Sutartinės than listeners without such expertise.

10.2.3 Affective priming

In affective priming tasks, participants are exposed to two stimuli in quick succession (the *prime*, here an interval, and the *target*, in this case a word). Participants classify the second (target) stimulus as positive or negative. Responses to affective priming tasks are generally accepted to be providing an implicit measure of attitudes to a stimulus (see e.g., Klauer, 1997; Wittenbrink & Schwarz, 2007). In particular, affective priming is well established as a measure of automatic responses to C/D e.g. Lahdelma, Armitage, and Eerola, 2022; Sollberger et al., 2003; Steinbeis and Koelsch, 2011. There is consensus in the music priming literature that targets are classified more quickly and accurately when the chord and the word are of the same valence (i.e. dissonant-negative or consonant-positive) compared to when the stimulus and the word are of opposing valence.

10.2.4 The present study

The present study aims to investigate how listeners expertise in a musical tradition with a high prevalence of acoustically rough intervals (i.e. the minor and major second in Lithuanian Sutartinės) influences their responses to these intervals compared to non-musicians and musicians trained in Western diatonic genres. In particular, we test whether there is a differential effect on automatic responses (which we predict are governed by psychoacoustic factors) and conscious evaluative responses (which we predict are governed by familiarity.) More specifically, we predict that automatic responses to an affective priming task are independent of musical background, whereas appraisal responses are be influenced by expertise in Sutartinės, i.e. by exposure to a particular musical culture in which minor and major second intervals are more prevalent. The influence of a musical culture rich in acoustic roughness on C/D perception is compared with the influence of expertise in Western tonal music, where acoustically rough intervals are used less frequently. We expect Sutatineės singers to rate the major second as more positive than do Western musicians, who we in turn expect to rate the major second as more positive than non-musicians due to higher familiarity. In the affective primining task, we expect that targets would be classified more quickly and more accurately when they are preceded by a congruent prime. However, we do not predict any group differences in this effect.

10.3 Method

10.3.1 Participants

Brysbaert and Stevens (2018) suggest that, for a mixed design with multiple items in each condition, 1600 readings per condition (i.e., 25 participants per group) are necessary to achieve a power of 0.8. To mitigate against attrition, we targeted a sample size of thirty per group. Following deletions (see data pre-treatment below), the Sutartinės group consisted of 24 members of Lithuanian Sutartinės singing groups (all female; mean age = 42.4, SD = 13.6; slightly below target sample size). The control group consisted 26 female native speakers of Lithuanian recruited via Prolific.co, an online crowd-sourcing platform designed especially for research purposes (mean age = 29.09, SD = 6.09). The Western musician and non-musician groups were also recruited via Prolific. The musician group were pre-filtered as having at least 5 years' experience in singing/playing an instrument. 28 musicians (all female, mean age = 46.5 years, SD = 15.2) and 31 non-musicians (all female, mean age = 37.2 years, SD = 11.5) completed the study. Participants received £3.25 (3.69) for participating in the study. The study received ethical approval from the Department of Music ethics committee, University of Anonymous for review (MUS-2021-02-04T10_57_12-ghth52). Informed consent was provided via an online checkbox.

10.3.2 Materials and Stimuli

To test automatic versus conscious appraisal responses to the intervals, we used two intervals: a perfect fifth (G_3 – D_4) (P5) and a flat major second (Bb_3 – C_4) (M2) tuned to 175 cents, consistent with the Sutartinės style which harnesses intervals slightly narrower than the equally-tempered major second, with a wide range of variations (Ambrazeviius, 2017). The M2-P5 combination has previously proven effective at generating affective priming effects (Armitage et al., 2021). The intervals were generated with *Ableton Live 9* (a music sequencer software), using the *Venus Symphonic Women's Choir* sample-based plug-in as the sound font. The stimuli were normalised by setting them to the same peak sone level. The duration of the stimuli were 800 ms with a 50 ms fade-out. The amplitudes and spectra for the two intervals are shown in Figure 10.3.1.

Target words in English and Lithuanian are listed in Table 10.3.1. The words were translated from English to Lithuanian by the fourth author; back translation to English by an independent native speaker of Lithuanian agreed with the original words. As in previous chapters, the English words were matched for length. However, in order to preserve the meanings across languages, the Lithuanian words were not matched for length. In particular, positive words were longer than nega-

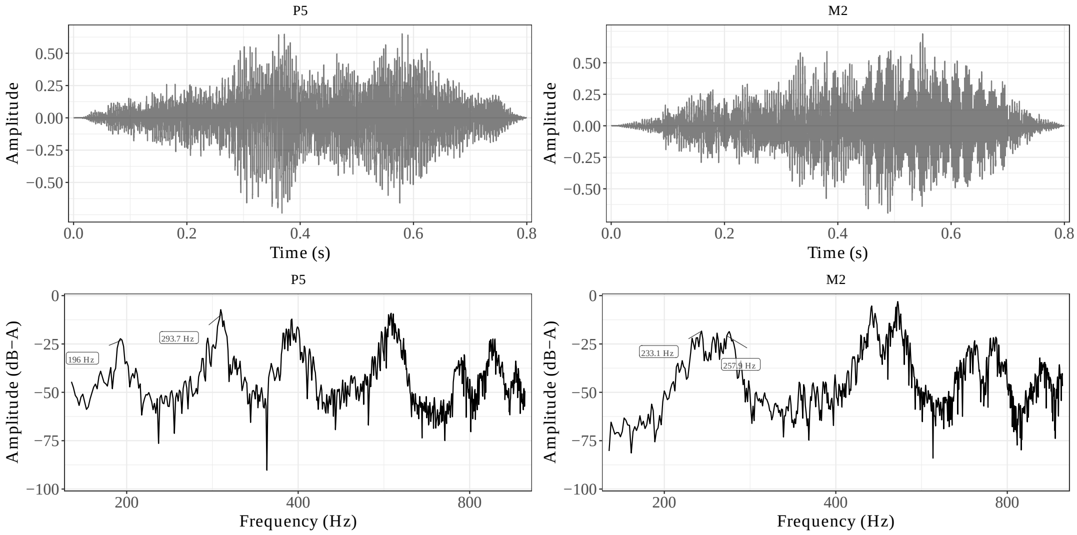


Figure 10.3.1: Waveforms and spectra for P5 and M2.

tive words in Lithuanian.[2] As each prime interval was paired with each target word, there were 32 possible prime-target combinations. Each combination could be either congruent (e.g. P5 - "lively", or M2 - "morgue") or incongruent (e.g. P5 - "dismal", M2 - "excite").

The priming experiments and the rating task were implemented online in PsyToolkit (Stoet, 2017). RT data collected via PsyToolkit and Prolific have previously been shown to be comparable to lab-based data collection (Armitage & Eerola, 2020).

10.3.3 Procedure

Participants were told that they were completing an experiment on the relationship between sounds and reading. Initially participants provided demographic information and details of their musical background and expertise. Next, in order to orient them to the musical style, participants heard a short extract from the Sutartinės *Atvazhiuok, mochiute* and *O kas prazhydo* which lasted a total of 2 minutes 45 sec-

[2]One potential consequence was that the positive words would be processed more slowly than negative words. However, RTs for (mean RT = 677 ms) words were faster than for negative words (mean RT = 700 ms), $t(23) = 3.17,\ p < 0.01,\ 95\%\ CI = [9.95, 47.40]$. This finding is consistent with previous chapters and the priming literature more generally.

Table 10.3.1: **Target Words in English and Lithuanian**

English	Lithuanian
climax	kulminacija
lively	gyvas
gentle	velnus
rest	ilstis
comfy	patogus
admire	avtis
payday	atlyginimo diena
relax	atsipalaiduoti
rabid	pasiutęs
hijack	pagrobti
coma	koma
arrest	aretuoti
flaccid	bevalis
morgue	morgas
coward	bailys
dismal	nirus

onds.

For the priming task, participants were presented with a series of target words which were to be categorised as either positive or negative. The task was restricted such that participants could only use laptop or desktop computers with a physical keyboard, i.e. it was not possible to complete the task via mobile phone or tablet. Participants were instructed to press the z key if the word had negative connotations or the m key if the word had positive connotations. Each word was preceded by a brief sounding of an interval. The timings for each item of the priming task was as follows. Initially, a fixation cross was presented on-screen. After 450 ms, the prime intervals sounded for 800 ms. 200 ms into the auditory presentation of the prime interval, the fixation cross was replaced with the target word, which remained onscreen for 2000 ms or until a key press.

The priming task consisted of a ten-item practice block, followed by two 64-item (i.e. two prime chords x 16 target words; each prime-target combination occurred twice) experimental block. During the practice block, participants received feedback after each item to indicate whether or not their answer was correct; no feedback was given during the experimental block. Items were presented in a random order.

For the rating task, participants rated the M2 and P5 intervals on a 7-point likert scale where 1 represented most negative and 7 represented most positive. Following the experiment, participants were debriefed as to the purpose of the study.

10.4 Results

10.4.1 Data Pre-Treatment and Statistical Analysis

Prior to analysis, five participants were removed from the data set in line with the exclusion criteria above. Incorrect responses to individual items were deleted from the data set, as were timeouts (i.e. instances where participants fail to classify the target word within 2000 ms). Each participants data was fitted to a Gamma distribution: responses to individual items that were faster than 250 ms or slower than the 95th percentile of the Gamma distribution were deleted from the data set.

Reaction time data were fitted to a generalised mixed model (Lo & Andrews, 2015), using the R library lme4 (Bates et al., 2015) with the fixed effects of group (Sutartinės performers vs controls) and congruence (congruent vs incongruent), with the random effect participant. The model was then subjected to a 4 (Group) $\times 2$ (Congruency: Congruent vs Incongruent) type III ANOVA. Standardised effect sizes are not reported for GLMM. All statistical analyses were carried out in R (R Core Team, 2020) at $\alpha = .05$. Where there are multiple comparisons, Bonferroni corrections have been used throughout.

10.4.2 Reaction Time Analysis

Mean reaction times (standard deviations) and accuracy rates (%) are given in Table 10.4.1.

	Sutartinės Singers	Lithuanian Non-musicians	Western Musicians	Western Non-musicians
Congruent	691 ms (147) 90.5%	654 ms (175) 87.8%	615 ms (115) 88.4%	638 ms (150.0) 87.0%
Incongruent	706 ms (143) 88.2 %	662 ms (176) 88.2%	621 ms (112) 85.6%	648 ms(151) 85.8%

Table 10.4.1: Mean (SD) reaction times and accuracy rates per condition

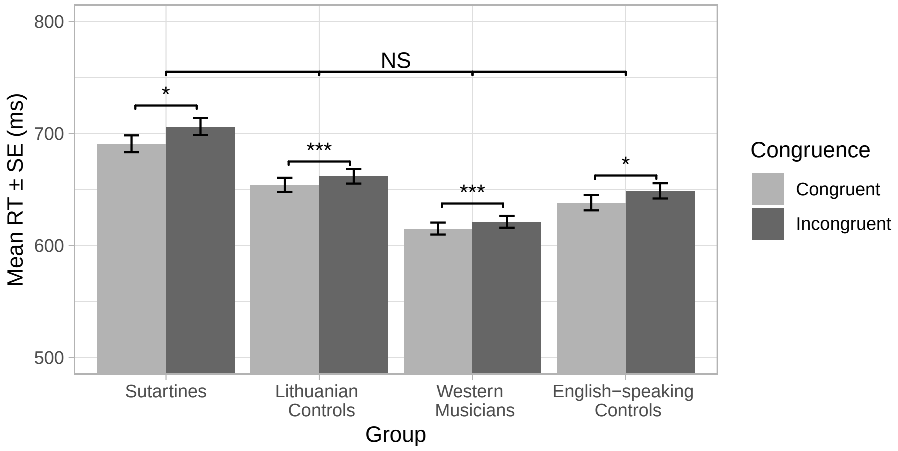


Figure 10.4.1: Mean and SE RT by congruency condition. Between groups differences were non-significant; for within groups differences, * denotes $p < .05$ *** denotes $p < .001$

The ANOVA[3] revealed a significant main effect of congruency, $\chi^2(1) = 32.79,\ p < .001$. Planned contrasts showed that RTs in the Congruent condition were significantly faster than those in the Incongruent condition, $z = 5.73,\ p < .0001,\ 95\% = [0.25, \infty]$. The congruency effect is represented graphically in Figure 10.4.1 The main effect of Group proved non-significant, $\chi^2(3) = 4.28,\ p = .24$, as did the interaction of Congruence and Group, $\chi^2(3) = 3.44,\ p = .33$.

[3]as an additional analysis, we analysed the data with a 4 (Group) x 2 (Prime: m2 vs P5) x 2 (Target Valence: Positive vs Negative) Type III Anova. Under this analysis, there were again no group differences present. In particular, we saw that for all groups, Positive targets were classified significantly more slowly when preceded by M2 compared to P5. This is supported by visual inspection of Figure , which suggests that inhibition is present in the M2-positive target condition for all groups of participants.

10.4.3 Accuracy Rate Analysis

Next, accuracy rates were subjected to a 2 (Group: Sutartinės vs Controls) ×2 (Congruence: Congruent vs Incongruent) type III ANOVA. The main effect of Congruence proved significant, $F(1, 105) = 10.26,\ p = .002$. Planned contrasts confirmed that ARs in congruent conditions were significantly higher than in incongruent conditions, $t(108) = 3.16\ p = .001,\ 95\%\ CI = [0.82, \infty],\ d = 0.31$ The main effect of Group proved non-significant $F(3, 105) = 0.82,\ p = .49$. The interaction of Congruence and Group was also non-significant, $F(3, 105) = 1.92,\ p = .13$.

10.4.4 Valence Ratings

Mean (SD) valence ratings for the stimuli are given in Table 10.4.2

	Sutartines Singers	Lithuanian Control	Western Musicians	Western Controls	Overall
M2	3.54 (1.74)	2.11 (1.03)	2.87 (1.38)	3.36 (1.50)	2.97 (1.51)
P5	5.67 (1.20)	4.73 (1.76)	5.14 (1.18)	5.19 (1.33)	5.15 (1.41)

Table 10.4.2: Mean (SD) Valence Ratings for Audio Stimuli

Finally, we considered the valence ratings for M2 and P5 from the Lithuanian and the English-speaking groups. The data were subject to a MANOVA with Pillai's trace, with the dependent variables M2 and P5 valence ratings and the independent variable of group.

We saw a significant main effect of Group on M2 ratings, $V(6, 212) = 2.9518,\ p = .008$. Subsequent univariate ANOVAs yielded a significant effect of group on M2 ratings, $F(3, 106) = 5.30,\ p = .002$. We used a planned contrasts approach to test our hypothesis that familiarity with M2 would lead to higher valence ratings, i.e., Sutartinės singers would rate M2 higher than other groups. This approach revealed significant differences in group ratings for M2 between Sutartinės singers and Lithuanian non-musicians, $t(106) = 3.62,\ p < .001$, and between Sutartinės singers and Western musicians $t(106) = 1.81,\ p = .04$. However, contrary to expectations, there was no evidence of a group difference between Sutartineės singers and Western non-musicians, $t(106) = 0.52,\ p = .30$.

10.5 Discussion

Automatic responses to musical intervals were found to follow a similar pattern, irrespective of musical culture or level of musical expertise. Congruency effects were found to be present in all four groups of participants, i.e., reaction times were shorter when the musical interval and the target word both had the same valence compared to when the musical interval and the target word had opposite valences and this finding is in line with recent research investigating the perception of C/D in intervals with an affective priming paradigm (Armitage et al., 2021; Steinbeis & Koelsch, 2011) targeting Western listeners. In particular, the Sutartinės singers did not differ in their responses from the other groups. This suggests that exposure to a musical culture where there is a high prevalence of parallel seconds does not influence automatic responses to those intervals. This in turn suggests that there is another factor that is responsible for the ability of the second to prime negative words, and that this is most likely an acoustic feature of the interval. Previous research (see Armitage et al., 2021; Lahdelma, Armitage, & Eerola, 2022; Steinbeis & Koelsch, 2011) has argued that roughness is the governing factor in automatic responses to intervals and more complex chords. In these instances, the negative valence allocated to 'rough' intervals such as seconds is thought to be a consequence of the inability of the basilar membrane to resolve the fundamental waves that are close in frequency – i.e., which lie within the same critical band – and is therefore not susceptible to modulation by musical experience.

In contrast to the automatic responses to the intervals, valence ratings did reveal group differences. As predicted, Sutartinės singers rated the flattened minor second as more positive than did their Western musician counterparts or participants in the

Lithuanian and Western non-musician groups. Although this presents an apparent contradiction with the result of the reaction time task, it seems likely that the valence ratings are based on a conscious mechanism that depends on acquired knowledge of particular musical sound-worlds, which allows for a more positive interpretation of the M2 interval, whereas this knowledge based understanding of the interval is not activated during the priming task. In such tasks the response is formulated at an earlier stage of processing, i.e., the evaluation as negative is in response to activity in the basilar membrane. This finding is in line with both neurological studies (Linnavalli et al., 2020; Pallesen et al., 2005) as well as affective priming studies (Armitage et al., 2021; Lahdelma, Armitage, & Eerola, 2022; Steinbeis & Koelsch, 2011) that have not found a difference in automatic reactions to consonant vs. dissonant (rough) pitch combinations among Western listeners, even if the conscious appraisal of such pitch combinations does differ according to musical expertise.

Although recent fieldwork across non-Western and Western populations has identified high amounts of roughness as a possible source for universal aversion (Lahdelma et al., 2021) in simultaneous pitch-combinations it has also been proposed that musical sounds are not rough enough to cause aversion that is determined by the properties of the human auditory system (McDermott et al., 2016). Indeed, the validity of the critical bandwidth theory has been questioned (see Maher, 1976) on the basis of findings that many non-Western musical cultures promote roughness for aesthetic ends in musical expression. Here we have demonstrated that these apparently contradictory findings may be reconciled if we separate the layers of acquired aesthetic, self-reported responses and automatic affective responses that are likely to be biologically determined due to constraints of the human auditory system.

Notably, the automatic aversion to the M2 interval found here is in line with McDermott et al.'s 2016 finding that the un-Westernised Tsimané people also had an aversion to major and minor seconds (i.e., the two intervals whose separate fundamentals fall within the same critical band) when presented diotically (simultaneous presentation to each ear). The Tsimané provide a particularly compelling case as in terms of exposure to simultaneous pitch combinations they can be considered as a *tabula rasa* – vertical harmony does not feature in the Tsimané musical culture. Hence, their responses are not likely to be the consequence of familiarity with either Western tonal harmony, or conversely, the rough harmonies prevalent in beat diaphony musical styles. Of additional interest here is the research of McPherson et al., 2020 conducted also on the Tsimané demonstrating that perceived fusion (the tendency for simultaneous sounds to blend perceptually or to be perceived as one sound) among the Tsimané was greater for the intervals of the octave, the fifth, and the fourth than for the dissonant intervals closest in size, just like in the case of Western listeners. Strikingly, fusion did not predict preferences in Tsimané participants, who did not indicate a preference for these fused (consonant) intervals. This reiterates the previous point that while the distinction between consonance/dissonance may have a biological basis in how the human auditory system works, this distinction does not entail a universal aesthetic response on a conscious level.

The present study addressed C/D in the context of Sutartinės singers and musicians in Western tonal traditions alongside non-musician controls from Lithuania and English-speaking countries. However, its scope is limited to intervals rather than more complex chords. It should also be noted that harmony perception is not usually limited to isolated intervals (or chords), i.e., harmony should be considered horizontally as well as vertically. As Persichetti (1961, p. 189) points out these two planes interact dynamically in actual music: "when melodies sound together chords are formed, and when chords follow each other melodic motion is involved...even the most isolated chord is full of melodic potential". It is clear that C/D is affected by context as well as musical style; for example what is restless and dissonant in common-practice tonality may be considered relaxed and consonant in jazz such as dissonant chords giving a sense of finality in jazz performances (Parncutt et al., 2019). Future research should also expand the scope of cultures considered in perception of C/D. In particular, practitioners of other roughness-rich musical traditions such as Ganga (Bosnia), Gamelan (Indonesia), Mijwiz (the Middle East) would provide fertile ground for exploring the relationship between C/D, biological factors, and culture. In addition, other musical features, such as timbre, whose perception could be influenced by both biological and cultural factors should be studied using a similar combination of experimental and self-report methods.

Acknowledgments

The authors gratefully acknowledge Akvilė Jadzgeviciute for assistance in translation.

Data Availability

All data, R scripts, PsyToolkit Scripts and experimental stimuli are available at https://osf.io/v8tkz/?view_only=aa93269e54a74262858eceb96e8fca73.

Appendix

10.A Graphical representation of 4 (Group) × 2 (Prime Valence) ×2 (Target Valence) design

Figure 10.A.1 below shows the pattern of RT results for the 4 (Group) × 2 (Prime Valence) × 2 (Target Valence). The critical interaction *Prime Valence* × *Target Valence* proved significant $t = 5.47,\ p < .001$. All two- and three-way interactions involving the factor *Group* were non-significant, suggesting the priming effect was constant between groups.

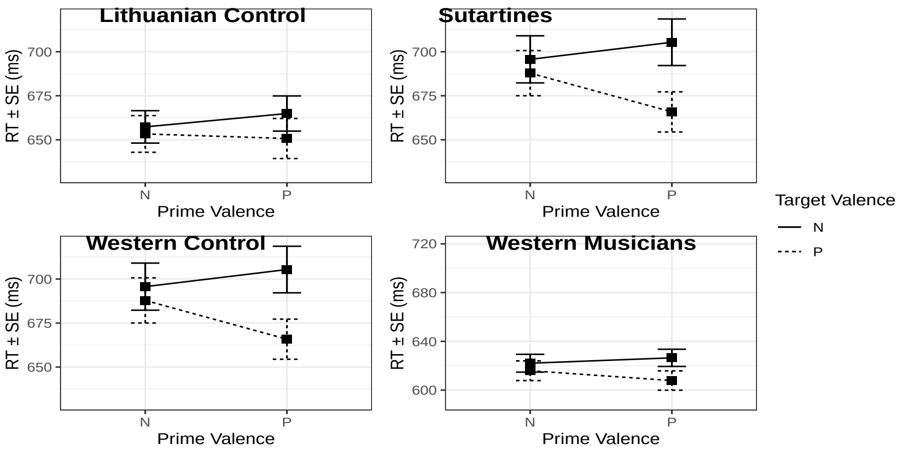


Figure 10.A.1: Prime Valence × Target Valence interaction by group

CHAPTER 11

General discussion and conclusions

Abstract

In the final chapter, the body of work produced in the thesis is summarised and the chapter explains how gaps in previous research have been addressed. Affective priming is discussed from cognitive and psychoacoustic perspectives, and the role of culture in the process is also considered. The main mechanism considered in the evaluation of target stimuli is top-down attention. Prime stimuli are considered in the context of bottom-up attention. It is argued that incongruent primes generate priming when they convey sufficient auditory information to disrupt the top-down process. The limitations of the research in the present thesis are discussed and suggestions for future research are put forward. In summary, the thesis explores how music and chord stimuli can influence judgements about emotional words, offering novel insights into the affective priming phenomenon, as well as demonstrating the efficacy of online data collection in music psychology research.

11.1 Introduction

This thesis set out to achieve three objectives. The first was to establish that affective priming could be reliably produced by music primes. The second objective was to gain insight into the cognitive processes involved in affective priming. The third and final objective was, once priming effects were established, to use affective priming as an indirect measure of consonance and dissonance by using reaction time measures as a way of probing acoustic properties of chord stimuli. These were addressed alongside a consideration of the scope and limitations of existing literature as well as setting reaction time methods in music psychology on a firmer analytical and empirical footing. Although reaction time, and specifically variants of priming, methods have been used in music psychology research for almost forty years, starting with Bharucha and Stoeckig (1986)s seminal priming papers, relatively little had been done with regard to validating the emotional properties of the stimuli and ensuring that the analysis strategy was robust. Moreover, much of the existing research has focused on establishing the existence of priming effects and not on probing how priming works on a perceptual or cognitive level; nor had the musical/acoustic properties of the music stimuli been explored thoroughly.

11.2 Summary of findings

The second objective of the present thesis was to explore the cognitive mechanisms responsible for allowing affective information in one modality - in this case music - to influence judgements about target information in another modality - written language. However, initially before it was possible to address how affective information was transferred between modalities, it was necessary to address the question of what affective information is transferred between modalities. Most previous affective priming studies had only considered how the valence of the prime stimulus could influence decisions about the valence of the target stimulus. Chapter 7 addresses the question of which dimension of emotion, valence or arousal, is transferred in affective priming.

Affective priming is typically explained in terms of a spreading activation account (Fazio et al., 1986) or in terms of Stroop-like interference. Chapter 8 attempts to address the question of which account provides the best explanation of priming. Indeed, attempting to unravel the mechanism responsible is an advance on previous music priming research: although the existence of the effects is well established, authors such as Costa (2013) and Sollberger et al. (2003) do not consider how the effect works. Chapter 8 suggests that priming effects are more parsimoniously explained by Stroop-like interference. This is borne out by drift-diffusion model results (Voss et al., 2013) which gave a non-decision time, t_0, that depended on the interaction of the target and prime valences. Moreover, given that the implication of the present thesis that Stroop interference is a more significant contributor to priming, and working memory is implicated as an important component of the priming mechanism. Indeed, this goes some way to explaining how priming is able to operate across modalities - although this could also be explained in terms of spreading activation: music primes activate concepts associated with the induced or perceived emotion in the semantic memory, and so target words are more easily and accurately categorised when the affective category they belong to has already been activated. However, the combined weight of the drift-diffusion model results and the support for the Attentional Control Theory (ACT) account of priming in participants high in trait anxiety suggests that Stroop-like interference is more plausible. It should be noted that the ACT result points to working memory activation in highly anxious participants; future research should address more directly working memory considerations in priming in a general sample. Although anxiety was considered as a candidate trait for influencing effective priming because of its clear links to fundamental cognitive mechanisms (i.e., a cognitivist perspective on personality), particularly emotion and attention, rather than being driven by an interest in traits per se, it is plausible that other traits may influence priming. Indeed, there is a small body of research that considers traits in domains other than music (e.g. neuroticism, empathy, psychopathy); a smaller number considers how traits and group differences – particularly alexithymia and Wilson's Syndrome (Goerlich et al., 2011; Lense et al., 2013) – influence music priming.

Having established the existence of priming effects and discussed the mechanisms responsible in the first half of the thesis, Chapters 9 and 10 go on to address the third objective: to use priming as an indirect measure in order to probe the acoustic properties of intervals in the context of consonance-dissonance research. This second part of the thesis builds on previous research by Costa; Sollberger et al. and Huziwara et al., and goes further than previous research by mapping priming effects to specific acoustic properties of intervals. These latter chapters assume that the word classification task depends on automatic responses to the target words. Chapter 9 demonstrated that stark differences in acoustic roughness are sufficient to produce priming effects in the cases of intervals, so m2 - P5 is sufficient to generate priming, whereas m3 - M3 is not. This is perhaps surprising given that major and minor triads are able to produce priming effects. Moreover, the musical excerpts used as primes in Part I of this thesis do not necessarily exhibit a high level of dissonance. It seems reasonable therefore to argue that there may be some combination of musical features that can contribute to the overall affective characteristics: highly rough chords or intervals may be sufficient to prime negative concepts or interfere with processing positive stimuli, but it is not necessary. In terms of the basic mechanisms that cause priming, this poses an interesting question: does arousal level amplify the priming effect? Chapter 9 argues that the high degree of roughness is linked to the inability of the basilar membrane to properly resolve minor and major seconds. This differentiates the rough intervals from other dissonant intervals such as the augmented fourth, which does not generate priming effects. One possibility is that the rough intervals generate a higher arousal level alongside the perceived negative valence, and that arousal level activates an urgency in the response to the negative stimulus. This is consistent with the research carried out by Hinojosa et al. (2009). However, this possibility has not been directly tested and remains an open question. Certainly, Chapter 7 demonstrates that the delineation between valence and arousal is not simply orthogonal.

Chapter 10 replicates the m2 - P5 priming found in Chapter 9. Importantly, priming effects caused by the m2-P5 combination were present across two groups of musical expertise – one, performers of Lithuanian beat diaphony, where the parallel squeezed seconds are used frequently for expressive purposes, and another, musicians who performed in Western diatonic traditions where vertical major and minor seconds are used more sparingly. It is striking to note that, whilst this automatic measure was present for musicians and and non-musicians in both cultures, there were significant differences in the conscious appraisals of the minor second: Sutartineės singers evaluated m2 as significantly more positive than did control groups or musicians with expertise in Western tonal idioms.

Furthermore, it is possible that the music excerpts and individual chords/intervals may not activate priming in exactly the same way. However, the results do offer apotential way forward to capture components of consonance and dissonance via an indirect method (priming) that does not rely on self-report. To that end priming should be more robust to observer and desirability effects than self-report methods.

In terms of methodological developments, Chapters 4 and 6 confirmed the feasibility of collecting reaction time data in music cognition research via online sources. In particular, Chapter 6 (published as Armitage & Eerola (2020)) demonstrated that priming effects can be detected online, whereas Chapter 4 surveys the broader context for online data collection in auditory research. The research presented in this thesis has taken place largely online. Although this was to some extent driven by necessity during the Covid-19 pandemic, it also demonstrates the priming effects are robust and can be detected outside of laboratory environments. However, it is possible that some results may have been clearer or more accurate with data collected in a lab setting. In particular, Chapters 9 and 10 may have benefitted from lab data as it depended on being confident in knowing what precisely participants heard. It should be noted that Chapter 6 found that the combination of PsyToolkit (Stoet, 2010, 2017) and Prolific replicated lab results, and that a small confirmatory sample of participants in controlled conditions replicated the same results as Chapter 9 with a stronger effect size. Thus, as well as the theoretical implications of the research in relation to priming mechanisms and using priming as an indirect measure of acoustic properties of stimuli in consonance-dissonance studies, the present research is also significant in moving forward the possibilities for using internet-mediated research in music cognition. Although several studies had used web-based questionnaires before, the studies published as part of this thesis demonstrate the feasibility of carrying out experimental music cognition research online. In particular, as well as the benefits of making access to large and diverse participant pools more accessible, online methods offer the potential to carry out music listening research in ecologically valid settings (i.e. at home, listening via a laptop or mobile device).

11.3 General discussion

Taken as a whole, the research in this thesis considers affective priming from cognitive, psychoacoustic and cultural perspectives. It is, however, important to assess the extent to which the three factors interact. Considering first of all the mechanisms involved in affective priming, Chapters 7 and 8 place attention as a central component of affective priming. Feature-specific attention allocation as proposed by Spruyt et al. (2012) was demonstrated, at least in part, to be capable of dictating which dimension of emotion, valence or arousal, was transferred in affective priming. Indeed, considered alongside other priming literature cite some here, it seems likely that top-down processes – simply the question of *should this stimulus be evaluated* *as positive or negative* – are central to the evaluative classification task. However,it is less clear whether or not priming effects in themselves are a result of the top-down process or a consequence of bottom-up interference in the response. Chapter 8 found that, in the case of affective environmental sounds, trait anxiety modulated responses to positive stimuli after negative primes. Much of the literature around anxiety considers susceptibility to bottom-up perception of threat-related stimuli. In particular, there is a strong consensus, reviewed in Mogg and Bradley (1999), suggesting that anxiety is linked to failures in top-down attentional control that allow bottom-up processes to overwhelm top-down processes. Following this line of reasoning, it seems likely that priming effects are a consequence of the bottom-up influence of the prime taking up attentional resources that should be deployed in fulfilling the top-down evaluative classification task. Tentatively, this is more consistent with an interference-based account of priming rather than a spreading activation account of priming. This is also consistent with ERP findings. For instance, Bartholow et al. (2009) argued in favour of a response conflict account of priming.

However, more recently, the role of top-down processes has become the focus of research activity in anxiety, for instance, LeDuke et al. (2023) and Sussman et al. (2016, 2020).

Affective priming is well-established in social cognition as an automatic measure of implicit attitudes to stimuli. This property of affective priming was exploited in Chapters 9 and 10 with regard to intervals (i.e., two-note chords). We saw in both Chapter 9 and Chapter 10 that acoustic roughness seems to be implicated in congruency effects, whereas stimuli such as minor thirds, which are typically considered as sad, but are low in acoustic roughness, were not capable of producing affective priming. It is important at this point to consider precisely what it is that makes acoustic roughness a determining factor in the presence or absence of affective priming. One possible explanation is that roughness is typically perceived as an auditory marker of threat and consequently acoustically rough stimuli activate bottom up processes that interfere with the top-down processes necessary for the evaluative classification task.

We also saw in Chapters 6, 7 and 8 that music clips can act as primes in both positive and negative conditions. Turning to negative stimuli in particular, it is important to note that in the case of intervals, negative low-arousal stimuli such as the minor 3rd or minor 6th are not associated with priming effects, whereas in the case of music clips, negative valence - low arousal stimuli (i.e., sad music) are associated with priming effects. Chapters 9 and 10 thus diverge somewhat from Chapters 8 and 9. Ratcliff, 1978 conceptualises speeded binary choice tasks as *evidence accumulation tasks*. One possible explanation is that intervals are sparse in musical information, whereas music clips contain multiple cues such as tonality, rhythm, timbre and register. Thus it seems plausible that information accumulation is the governing factor in whether priming is present – i.e., the question is does a given stimulus provide enough information in a given time interval to interfere with the judgement. Indeed, this is not incompatible with the results for acoustically rough intervals presented in 8 and 9 in that acoustically rough vocalisations are considered by some authors (Arnal et al., 2015, 2019) to be information-rich.

Information accumulation may also be implicated in the question of stimulus onset asynchrony. As reported in Chapter 3 – for instance, by comparing Ragozzine (2011) with for example Sollberger et al. (2003), chordal stimuli seem more effective at creating priming effects at shorter SOA. However, the present research (Chapters 5, 6 and 7) and Scherer and Larsen, 2011 have detected clear priming effects at SOA = 450 ms using information-rich stimuli, an SOA longer than many other studies. Certainly, one possible explanation for the presence of priming effects of such a (relatively) long SOA is the amount of auditory information contained in the stimulus. Thus, it seems plausible that, at 250 ms, most intervals, with the exception of minor and major seconds, do not convey sufficient auditory information to generate congruency effects. However, Costa (2013) and Sollberger et al. (2003) and Huziwara et al. (2023) found that triads and four-note chords were sufficient to generate priming. Indeed, in a study related to the present thesis, Lahdelma, Armitage, and Eerola (2022) tested directly the role of numerosity (i.e., the number of single tones in a chord) in priming, comparing two- and four-note chords and found that four-note chords were able to generate priming effects whereas intervals were not. Lahdelma et al. ascribed the possible causes of this difference to either information (i.e., four notes can be thought of as six distinct intervals) or roughness, which, in one sense, can be thought of as a specific kind of threat-related auditory information (Arnal et al., 2019). To date, the studies that have considered the size of affective priming effects as a function of SOA have considered this as a univariate problem – i.e., by considering some measure of the size of the priming effect as the dependent variable and SOA as the independent variable. However, it may be more informative to consider some measure of the amount of auditory information contained in the prime stimulus as the independent variable.

Thus far, the discussion has focused on how cognition and stimulus features, whether psychoacoustic or information density, interact to influence affective priming. At this point, it is important to consider the role of culture and whether and how culture can influence priming and probe the extent to which it interacts with cognitive and psychoacoustic factors. In Chapter 10, we saw that culture did not influence results in an affective priming task when the negative prime was a 'squeezed' second. In particular, we probed the difference between musicians in cultures where there is a high prevalence of parallel seconds and musicians experienced only in Western tonal idioms. Whilst priming – an automatic measure – did not reveal a group difference, we did see a group difference in conscious appraisal of the 'squeezed' second: it was rated as more positive by the Sutartineės singers. This raises an important question: is priming fundamentally independent of culture and learning, or is there a specific property of the stimuli (i.e., the 'squeezed' second vs perfect fifth) that makes it impervious to the influence of culture? It seems likely that the second is perceived as negative, at least during the short time frame used for a priming experiment as a consequence of its acoustic roughness – i.e., its effect on the basilar membrane. Turning to other priming studies, we see that Sollberger et al., 2003 and Costa, 2013 found priming with major and minor chords. While these chords have strong associations with happiness-sadness/positivity-negativity in Western cultures, they are not intrinsically different from an acoustic standpoint – both consist of a stacked major third and a minor third and are not substantially different from one another in terms of roughness and harmonicity. Indeed, cross-cultural research (Athanasopoulos et al., 2021) suggests that, in some cultures, the positive-negative association for major-minor is reversed so that minor chords are considered more positive than major chords. Consequently, it seems plausible that, in the absence of a strong psychoacoustic marker of 'negativity' such as acoustic roughness, in cultures where the major-minor association is reversed that major chords could function as negative primes and minor chords could function as positive primes. Furthermore, the extent to which the positive and negative connotations of stimuli are learnable – i.e., is it possible to condition listeners to consider particular sounds as positive or negative to such an extent that the sounds can then function as positive and negative primes – is an open question (and is indeed the subject of ongoing research by the present author and colleagues).

Overall, when considering the relationship between cognitive, psychoacoustic, and cultural factors that influence affective priming, it seems that a useful formulation may be to reconsider its genesis in social cognition and consider affective priming as an implicit measure of attitudes. In a top-down task – evaluative classification of targets – the prime stimuli engage bottom-up mechanisms. The extent to which they are able to do this depends on a participant's attitude to the prime: these attitudes may be formed by acoustic features that are potentially invariant, or by cultural influences. Where, as in the case of Sutartinės singers the acoustic and cultural evaluations differ, it seems to be that, in this case at least, nature has a stronger influence than nurture.

11.3.1 Recommendations for future research

In light of the interaction of the different cognitive mechanisms and stimulus features that contribute to priming, there are several directions that could be followed in future research. These include exploring the modality, the role of attention, individual differences, and methodological details. Chapters 7 and 8 suggest that both task specific attention allocation and attentional control are implicated in affective priming. However, there is a much broader research landscape in attention - much of which has not been applied in the context of priming. Certainly, future research should consider further whether there is a role for different attentional frameworks beyond task specific attention allocation and whether and how musical primes can orient attention to stimuli in a different modality. This seems plausible given the group differences in anxious vs low anxious participants discussed above. Again, the role of cross-modality is important in probing how auditory attention can be used to orient a listeners attention towards or away from a valenced target.

Future research should consider the role of other personality traits, such as empathy, which has been shown to be linked to susceptibility to music-induced emotion (Eerola et al., 2016), in modulating affective priming effects with musical stimuli. Where a cognitive model of a trait is well-developed and that trait is found to be related to priming, it has the potential to shed light on the cognitive mechanisms implicated in priming.

In addition, more research is needed to more precisely specify the role of modality in affective priming. Current research in music priming has been concerned largely with demonstrating that music primes can influence responses to targets in other modalities, in particular words. There has been limited discussion around the stimulus properties of both primes (particularly chords, where numerosity, timbre, modality, consonance-dissonance and pitch have all been considered) or targets (words and pictures). However, the mechanism by which this works has received scant attention. Both the Stroop and spreading activation accounts of priming have the potential to provide an explanation of how a stimulus in one modality can influence responses to a stimulus in another modality. Although Chapter 8 provides early evidence that working memory is at least in part responsible, it does not preclude completely the possibility of a spreading activation account which allows for the possibility of cross-modal corresponsences in semantic memory. Outside of the priming literature, there is promising discussion of cross-modal correspondances between music, vision and haptics. A full account of the cognitive mechanism of priming must take into account the cross-modal nature of the effect.

Chapter 3 suggested that SOA is a critical factor in whether or not priming effects are present. As discussed above, SOA alone may provide too simplistic a predictor of whether priming effects will be present. Future research should introduce an integrated measure of SOA and stimulus complexity and strength as a measure of auditory/musical information in order to predict the size of the priming effect.

Chapter 10 considers the role of culture in affective priming. It is argued that culture does not influence automatic responses to the major second (i.e., the priming effect is comparable in Sutartinės singers and controls) but does influence conscious appraisal (i.e., Sutartinės singers rate the major second as more positive than do controls). This is consistent with Tekman and Bharucha's (1998) findings regarding psychoacoustic versus cognitive mechanisms of harmonic priming. To probe these issues further, future research should attempt to delineate when cultural familiarity with a musical style, tuning system or stimulus set does or does not influence priming. Finally, alongside the role of long-term knowledge (i.e., culture) the role of short-term learning should be investigated. To this end, the present author is collaborating in a series of experiments to test the extent to which neutral sounds can be assigned positive or negative connotations and subject to a learning paradigm (e.g., by gamification) that uses the sounds as positive or negative primes.

Limitations

It should be noted that the research has a number of limitations. The majority of the research was carried out with native English speakers. Moreover, the music excerpts in Chapters 5, 6, 7 and 8 are taken from the Western musical tradition. In recent years, there has been much discussion of ethnocentricity, or over-reliance on on WEIRD samples (de Oliveira & Baggs, 2023), in psychological science. However, one experiment was carried out in Lithuanian with performers in a Lithuanian folk singing tradition and non-musician controls. Nevertheless, stronger evidence for generalisability, with participants from a wider pool of linguistic and musical backgrounds is necessary before the phenomena discussed can be declared universal. It is also possible that there may be a degree of self-selection in the participant pool who subscribe to Prolific. However, participant pools for many priming studies (e.g., Costa, 2013; Huziwara et al., 2023; Ragozzine, 2011; Sollberger et al., 2003; Tenderini et al., 2022; Zhou et al., 2019) use students as participant pool so the studies in the present thesis arguably may use a more diverse participant pool than the majority of existing music priming studies (see e.g., Peer et al., 2022, for discussion of diversity in crowdsourced vs convenience student samples). Although the stimuli for Chapters 5, 6 and 7 were taken from existing corpora of emotional music (Eerola & Vuoskoski, 2011; Västfjäll, 2001), it is assumed that it is the overall emotional properties of the music considered as a sort of gestalt that are responsible for the priming. It is possible that some specific musical feature is primarily responsible, such as mode, consonance, dissonance, timbre or event density.

Conclusions

Affective priming effects are well-established in a range of domains in psychology, spanning cognitive psychology, linguistics, social psychology and personality. In music settings, affective priming has been demonstrated by several authors in the last 20 years, e.g. Costa, 2013; Goerlich et al., 2011; Lense et al., 2013; Sollberger et al., 2003. However, relatively little attention has been paid to unraveling the cognitive mechanism responsible for priming effects, or to the properties of stimuli that generate priming. Chapters 7 and 8 consider which components of emotion, valence or arousal, can be transferred via affective priming before using models of anxiety to probe the mechanism for cross-modal priming.

Chapter 7 demonstrates that both valence and arousal can be transferred via affective priming. Moreover, they demonstrated a potential role of task-specific attention allocation, although the result was weaker than that found by Spruyt et al. (2012) Taken together with the role of anxiety, this suggests that attention plays a significant role in affective priming, despite the relative lack of attention received in the priming literature. A significant exception to this is the series of studies by Spruyt and colleagues.

Chapters 9 and 10 considered priming as an indirect measure of consonance and dissonance in musical intervals. Chapter 9 considered the psychoacoustic features roughness and harmonicity as predictors of affective priming. Although much of the present literature frames consonance and dissonance in terms of harmonicity, contrasts in harmonicity were not found to be predictors of consonance and dissonance whereas contrasts in roughness did predict priming.

The empirical section of the thesis concludes with an investigation into whether culture can influence automatic responses to musical intervals. Expert performers in Lithuanian Sutartinės completed and intervals priming study using a flattened major second as the negative prime compared to a major fifth as the positive prime. A control group of Lithuanian non-musicians and two groups of English-speaking participants (Western musicians vs non musicians) also completed the task. All groups showed priming effects, which was consistent with predictions based on the assertion that intervals of a major or minor second are considered acoustically 'rough' - i.e.

their fundamental frequencies are sufficiently close together that the basilar membrane cannot effectively distinguish between them. This suggests that an automatic aversion to such intervals is universal. However, on an explicit rating task, Sutartinės singers rated the second as more positive than did the other groups, suggesting that cultural familiarity with an interval influences conscious evaluations. Although current research on consonance and dissonance focuses on interaction between acoustics and culture, the difference between automatic and conscious evaluations of musical sounds is something that has received relatively little attention in the research literature to date.

Overall, the present thesis has examined music-induced affective priming from cognitive, psychoacoustic and cultural standpoints. In terms of contribution to understanding music priming, the contribution is threefold:

1. Chapters 7 and 8 address in greater depth the cognitive components of affective priming with music compared to previous research. In particular, the role of attention and the interaction of top-down vs bottom-up processes have been given greater consideration than has been the case in previous research.

2. Chapters 9 and 10 address the role of acoustic roughness, compared to harmonicity, in affective priming. Although roughness was considered as a potential driver of congruency effects in previous literature, (e.g., Steinbeis & Koelsch, 2011), this had not been probed systematically, and nor had it been orthogonally manipulated with other candidate features, such as harmonicity.

3. Chapter 10 discusses the role of culture in affective priming, and the extent to which affective priming, compared to self-report methods, is malleable or invariant under exposure to an acoustic phenomenon that is thought by some authors to be a biological universal. This contributes both to the discussion of affective priming and to the broader discussion of consonance and dissonance.

Additionally, the research has contributed methodologically to the field by solidifying the role of online data collection for behavioural (compared to self-report) methods.

Bibliography

Aarden, B. J. (2003). *Dynamic melodic expectancy* [Doctoral dissertation, The Ohio State University].

Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PloS One*, *12*(3), e0173392. https://doi.org/10.1371/journal.pone.0173392

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01535-9.

Ambrazeviius, R. (2017). Dissonance/roughness and tonality perception in Lithuanian traditional Schwebungsdiaphonie. *Journal of Interdisciplinary Music Studies*, *8*, 39–53.

Ambrazeviius, R. (2021). Migration of song genres: Two typical lithuanian cases. *European Journal of Musicology*, *20*(1), 105–132. https://doi.org/10.5450/EJM.20.1.2021.105

Ambrazeviius, R., Budrys, R., & Vinevska, I. (2015). *Scales in lithuanian traditional music: Acoustics, cognition, and contexts*. Kaunas University of Technology.

Ambrazeviius, R., & Winiewska, I. (2009). Tonal hierarchies in Sutartins. *Journal of Interdisciplinary Music Studies*, *3*(1/2), 45–55.

Ando, S., Kida, N., & Oda, S. (2002). Practice effects on reaction time for peripheral and central visual fields. *Perceptual and Motor Skills*, *95*(3), 747–751. https://doi.org/10.2466/pms.2002.95.3.747

Anglada-Tort, M., Harrison, P. M. C., & Jacoby, N. (2021). Repp: A robust cross-platform solution for online sensorimotor synchronization experiments. *bioRxiv*. https://doi.org/10.3758/s13428-021-01722-2

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, 1–19. https://doi.org/10.3758/s13428-020-01501-5

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Armitage, J., Lahdelma, I., & Eerola, T. (2021). Automatic responses to musical intervals: Contrasts in acoustic roughness predict affective priming in western listeners. *Journal of the Acoustical Society of America*, *150*(551). https://doi.org/https://doi.org/10.1121/10.0005623

Armitage, J., & Eerola, T. (2020). Reaction time data in music cognition: Comparison of pilot data from lab, crowdsourced, and convenience web samples. *Frontiers in Psychology*, 2883. https://doi.org/10.3389/fpsyg.2019.02883

Armitage, J., & Eerola, T. (2022). Cross-modal transfer of valence or arousal from music to word targets in affective priming? *Auditory Perception & Cognition*, *5*(3-4), 192–210. https://doi.org/10.1080/25742442.2022.2087451

Armitage, J., Lahdelma, I., Eerola, T., & Ambrazeviius, R. (2023). Culture influences conscious appraisal of, but not automatic aversion to, acoustically rough musical intervals. *PLoS ONE*, *18*. https://doi.org/https://doi.org/10.1371/journal.pone.0294645

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051–2056. https://doi.org/10.1016/j.cub.2015.06.043

Arnal, L. H., Kleinschmidt, A., Spinelli, L., Giraud, A.-L., & Mégevand, P. (2019). The rough sound of salience enhances aversion through neural synchronisation. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-11626-7

Athanasopoulos, G., Eerola, T., Lahdelma, I., & Kaliakatsos-Papakostas, M. (2021). Harmonic organisation conveys both universal and culture-specific cues for emotional expression in music. *PloS One*, *16*(1), e0244964. https://doi.org/https://doi.org/10.1371/journal.pone.0244964

Aubé, W., Angulo-Perkins, A., Peretz, I., Concha, L., & Armony, J. L. (2015). Fear across the senses: Brain responses to music, vocalizations and facial expressions. *Social cognitive and affective neuroscience*, *10*(3), 399–407. https://doi.org/https://doi.org/10.1093/scan/nsu067

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British journal of psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bakker, D. R., & Martin, F. H. (2015). Musical chords and emotion: Major and minor triads are processed for emotion. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 15–31. https://doi.org/10.3758/s13415-014-0309-4

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, (22), 153–160. https://doi.org/10.1136/ebmental-2019-300117

Bannister, S. (2020). A vigilance explanation of musical chills? Effects of loudness and brightness manipulations. *Music & Science*, *3*, 2059204320915654.

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PloS One*, *10*(9), 1–29. https://doi.org/10.1371/journal.pone.0136675

Bargh, J. A., Lombardi, W. J., & Higgins, E. T. (1988). Automaticity of chronically accessible constructs in person× situation effects on person perception: It's just a matter of time. *Journal of Personality and Social Psychology*, *55*(4), 599. https://doi.org/10.1037/0022-3514.55.4.599

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & Van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, *133*(1), 1. https://doi.org/10.1037/0033-2909.133.1.1

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using qualtrics. *Behavior Research Methods*, *47*(4), 918–929. https://doi.org/https://doi.org/10.3758/s13428-014-0530-7

Barsalou, L. W. (2014). Cognitive psychology: An overview for cognitive scientists.

Bartholow, B. D., Riordan, M. A., Saults, J. S., & Lust, S. A. (2009). Psychophysiological evidence of response conflict and strategic control of responses in affective priming. *Journal of Experimental Social Psychology*, *45*(4), 655–666. https://doi.org/10.1016/j.jesp.2009.02.015

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, A. T., & Clark, D. A. (1988). Anxiety and depression: An information processing perspective. *Anxiety research*, *1*(1), 23–36. https://doi.org/10.1080/10615808808248218

Beck, A. T., & Clark, D. A. (1997). An information processing model of anxiety: Automatic and strategic processes. *Behaviour research and therapy*, *35*(1), 49–58. https://doi.org/10.1016/S0005-7967(96)00069-1

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*(6), 893. https://doi.org/10.1037/0022-006X.56.6.893

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800. https://doi.org/10.3758/s13428-011-0081-0

Bellec, G., Elowsson, A., Friberg, A., Wolff, D., & Weyde, T. (2013). A social network integrated game experiment to relate tapping to speed perception and explore rhythm reproduction. *Proceedings of the Sound and Music Computing Conference*, 19–26.

Bentz, D., & Schiller, D. (2015). Threat processing: Models and mechanisms. *Wiley interdisciplinary reviews: cognitive science*, *6*(5), 427–439. https://doi.org/10.1002/wcs.1353

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bermudez, P., & Zatorre, R. J. (2009). A distribution of absolute pitch ability as revealed by computerized testing. *Music Perception: An Interdisciplinary Journal*, *27*(2), 89–101. https://doi.org/10.1525/mp.2009.27.2.89

Berner, M. P., & Maier, M. A. (2004). The direction of affective priming as a function of trait anxiety when naming target words with regular and irregular pronunciation. *Experimental Psychology*, *51*(3), 180–190. https://doi.org/10.1027/1618-3169.51.3.180

Bharucha, J. J., & Stoeckig, K. (1986). Reaction time and musical expectancy: Priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 403–410. https://doi.org/10.1037/0096-1523.12.4.403

Bharucha, J. J., & Stoeckig, K. (1987). Priming of chords: Spreading activation or overlapping frequency spectra? *Perception & Psychophysics*, *41*(6), 519–524. https://doi.org/10.3758/BF03210486

Bhatti, S. S., Gao, X., & Chen, G. (2020). General framework, opportunities and challenges for crowdsourcing techniques: A comprehensive survey. *Journal of Systems and Software*, *167*, 110611. https://doi.org/10.1016/j.jss.2020.110611

Bigand, E., Filipic, S., & Lalitte, P. (2005). The time course of emotional responses to music. *Annals of the New York Academy of Sciences*, *1060*, 429–437. https://doi.org/10.1196/annals.1360.036

Bigand, E., McAdams, S., & Forźt, S. (2000). Divided attention in music. *International Journal of Psychology*, *35*(6), 270–278. https://doi.org/https://doi.org/10.1080/002075900750047987

Bishop, D. T., Karageorghis, C. I., & Kinrade, N. P. (2009). Effects of musically-induced emotions on choice reaction time performance. *The Sport Psychologist*, *23*(1), 59–76. https://doi.org/10.1123/tsp.23.1.59

Blacking, J. (1995). *Music, culture, and experience: Selected papers of John Blacking*. University of Chicago Press.

Blair, K. S., Richell, R. A., Mitchell, D. G., Leonard, A., Morton, J., & Blair, R. J. (2006). They know the words, but not the music: Affective and semantic priming in individuals with psychopathy. *Biological Psychology*, *73*(2), 114–123. https://doi.org/10.1016/j.biopsycho.2005.12.006

Blaison, C., Imhoff, R., Hühnel, I., Hess, U., & Banse, R. (2012). The affect misattribution procedure: Hot or not? *Emotion*, *12*(2), 403. https://doi.org/10.1037/a0026907

Boakes, J. A. (2010). *The role of specific emotions in affective priming effects*. University of Western Australia.

Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the united states and india: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*, *8*(2), 232–250. https://doi.org/10.1017/psrm.2018.28

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese*, *193*, 1509–1534. https://doi.org/10.1093/oso/9780198866282.003.0009

Bowling, D. L., Hoeschele, M., Gill, K. Z., & Fitch, W. T. (2017). The nature and nurture of musical consonance. *Music Perception: An Interdisciplinary Journal*, *35*(1), 118–121. https://doi.org/10.1525/mp.2017.35.1.118

Bowling, D. L., & Purves, D. (2015). A biological rationale for musical consonance. *Proceedings of the National Academy of Sciences*, *112*(36), 11155–11160. https://doi.org/10.1073/pnas.1505768112

Bowling, D. L., Purves, D., & Gill, K. Z. (2018). Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences*, *115*(1), 216–221. https://doi.org/10.1073/pnas.1713206115

Bradley, M. M., & Lang, P. J. (2007). The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3*.

Bradshaw, A., & McGettigan, C. (2021). Synchronised speech and speech motor control: Convergence in voice fundamental frequency during choral speech. https://doi.org/10.31234/osf.io/9hc34

Bradt, J., Dileo, C., & Shim, M. (2013). Music interventions for preoperative anxiety. *Cochrane Database of Systematic Reviews*, (6). https://doi.org/https://doi.org/10.1002/14651858.CD006908.pub2

Brandl, R. M. (1989). Die schwebungs-diaphonie–aus musikethnologischer und systematisch-musikwissenschaftlicher sicht. *Südosteuropa-Studien. Bd*, *40*, 51–67.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the mcgurk effect? *PloS One*, *13*(11), 1–20. https://doi.org/10.1371/journal.pone.0207160

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.72

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Burgoyne, J. A., Bountouridis, D., van Balen, J., & Honing, H. (2013). Hooked: A game for discovering what makes music catchy. *Proceedings of the 14th Society of Music Information Retrieval Conference (ISMIR).*

Cańs, J. J. (1990). Associative strength effects in the lexical decision task. *The Quarterly Journal of Experimental Psychology*, *42*(1), 121–145. https://doi.org/https://doi.org/10.1080/14640749008401211

Carcagno, S., Lakhani, S., & Plack, C. J. (2019). Consonance perception beyond the traditional existence region of pitch. *The Journal of the Acoustical Society of America*, *146*(4), 2279–2290. https://doi.org/10.1121/1.5127845

Carr, T. H., & Dagenbach, D. (1990). Semantic priming and repetition priming from masked words: Evidence for a center-surround attentional mechanism in perceptual recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 341. https://doi.org/10.1037/0278-7393.16.2.341

Carroll, N. C., & Young, A. W. (2005). Priming of emotion recognition. *The Quarterly Journal of Experimental Psychology Section A*, *58*(7), 1173–1197. https://doi.org/10.1080/02724980443000539

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, *7*(2), 2158244017712774.

Casler, K., Bickel, L., & 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), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Castaneda, J. O., & Segerstrom, S. C. (2004). Effect of stimulus type and worry on physiological response to fear. *Journal of anxiety disorders*, *18*(6), 809–823. https://doi.org/https://doi.org/10.1016/j.janxdis.2003.10.003

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038.

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cohen, H. F. (1984). *Quantifying music: The science of music at the first stage of scientific revolution 1580–1650* (Vol. 23). Springer Science & Business Media.

Costa, M. (2013). Effects of mode, consonance, and register in visual and word-evaluation affective priming experiments. *Psychology of Music*, *41*(6), 713–728. https://doi.org/10.1177/0305735612446536

Costa, M., Ricci Bitti, P. E., & Bonfiglioli, L. (2000). Psychological connotations of harmonic musical intervals. *Psychology of Music*, *28*(1), 4–22. https://doi.org/10.1177/0305735600281002

Cousineau, M., McDermott, J. H., & Peretz, I. (2012). The basis of musical consonance as revealed by congenital amusia. *Proceedings of the National Academy of Sciences*, *109*(48), 19858–19863. https://doi.org/10.1073/pnas.1207989109

Cuccolo, K., Irgens, M. S., Zlokovich, M. S., Grahe, J., & Edlund, J. E. (2021). What crowdsourcing can offer to cross-cultural psychological science. *Cross-Cultural Research*, *55*(1), 3–28. https://doi.org/10.1177/1069397120950628

Curtis, M. E., & Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, *10*(3), 335. https://doi.org/10.1037/a0017928

Dannlowski, U., Kersting, A., Lalee-Mentzel, J., Donges, U.-S., Arolt, V., & Suslow, T. (2006). Subliminal affective priming in clinical depression and comorbid anxiety: a longitudinal investigation. *Psychiatry Research*, *143*(1), 63–75. https://doi.org/10.1016/j.psychres.2005.08.022

Davidson, R. J., Sherer, K. R., & Goldsmith, H. H. (2009). *Handbook of affective sciences*. Oxford University Press. https://doi.org/10.1093/oso/9780195126013.001.0001

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin*, *135*(3), 347.

De Luca, R., & Botelho, D. (2020). Olfactory priming on consumer categorization, recall, and choice. *Psychology & Marketing*, *37*(8), 1101–1117. https://doi.org/10.1002/mar.21342

Degner, J., Doycheva, C., & Wentura, D. (2012). It matters how much you talk: On the automaticity of affective connotations of first and second language words. *Bilingualism: Language and Cognition*, *15*(1), 181–189. https://doi.org/10.1017/S1366728911000095

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a web browser? Comparing response times collected with javascript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/https://doi.org/10.3758/s13428-015-0567-2

Deličge, I., & Sloboda, J. A. (2004). *Perception and cognition of music*. Psychology Press.

Delignette-Muller, M. L., & Dutang, C. (2015). Fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(4), 1–34. http://www.jstatsoft.org/v64/i04/

de Oliveira, G. S., & Baggs, E. (2023). *Psychology's weird problems*. Cambridge University Press.

Dignath, D., Janczyk, M., & Eder, A. B. (2017). Phasic valence and arousal do not influence post-conflict adjustments in the simon task. *Acta Psychologica*, *174*, 31–39. https://doi.org/10.1016/j.actpsy.2017.01.004

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? screening mechanical turk workers. *Proceedings of the SIGCHI conference on human factors in computing systems*, 2399–2402.

Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, *13*(6), 513–519. https://doi.org/10.1111/1467-9280.00490

Duffy, S., & Pearce, M. (2018). What makes rhythms hard to perform? An investigation using steve reich's clapping music. *PloS One*, *13*(10), e0205847. https://doi.org/10.1371/journal.pone.0205847

Eerola, T., & Vuoskoski, J. K. (2012). A review of music and emotion studies: Approaches, emotion models and stimuli. *Music Perception*, *30*(3), 307–340. https://doi.org/10.1525/mp.2012.30.3.307

Eerola, T., Armitage, J., Lavan, N., & Knight, S. (2021). Online data collection in auditory perception and cognition research: Recruitment, testing, data quality and ethical considerations. *Auditory Perception & Cognition*, *4*(3-4), 251–280. https://doi.org/10.1080/25742442.2021.2007718

Eerola, T., & Lahdelma, I. (2021). The anatomy of consonance/dissonance: Evaluating acoustic and cultural predictors across multiple datasets with chords. *Music & Science*, *4*, 1–19. https://doi.org/10.1177/20592043211030471

Eerola, T., & Lahdelma, I. (2022). Register impacts perceptual consonance through roughness and sharpness. *Psychonomic Bulletin & Review*, *29*(3), 800–808. https://doi.org/10.3758/s13423-021-02033-5

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49. https://doi.org/10.1177/0305735610362821

Eerola, T., Vuoskoski, J. K., & Kautiainen, H. (2016). Being moved by unfamiliar sad music is associated with high empathy. *Frontiers in Psychology*, *7*(SEP). https://doi.org/10.3389/fpsyg.2016.01176

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336. https://doi.org/10.1037/1528-3542.7.2.336

Faria, M., Björnmalm, M., Thurecht, K. J., Kent, S. J., Parton, R. G., Kavallaris, M., Johnston, A. P., Gooding, J. J., Corrie, S. R., Boyd, B. J., et al. (2018). Minimum information reporting in bio–nano experimental literature. *Nature Nanotechnology*, *13*(9), 777–785. https://doi.org/10.1038/s41565-018-0246-4

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and models* (Vol. 22). Springer Science & Business Media.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion*, *15*(2), 115–141. https://doi.org/10.1080/02699930125908

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229. https://doi.org/10.1037/0022-3514.50.2.229

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not english. *Personality and Individual Differences*, *75*, 47–52. https://doi.org/10.1016/j.paid.2014.11.017

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. Labvanced: A unified javascript framework for online studies. In: In *International conference on computational social science (cologne)*. 2017.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, *12*(4), 377–430. https://doi.org/10.1121/1.1915637

Forero, D. A., Lopez-Leon, S., Gonzļlez-Giraldo, Y., & Bagos, P. G. (2019). Ten simple rules for carrying out and writing meta-analyses. *PLoS Computational Biology*, *15*(5), e1006922. https://doi.org/10.1371/journal.pcbi.1006922

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991. https://doi.org/10.3758/s13423-012-0322-y

Freytag, P., Bluemke, M., & Fiedler, K. (2011). An adaptive-learning approach to affect regulation: Strategic influences on evaluative priming. *Cognition and Emotion*, *25*(3), 426–439. https://doi.org/10.1080/02699931.2010.537081

Friedman, R. S., Kowalewski, D. A., Vuvan, D. T., & Neill, W. T. (2021). Consonance preferences within an unconventional tuning system. *Music Perception: An Interdisciplinary Journal*, *38*(3), 313–330. https://doi.org/10.1525/mp.2021.38.3.313

Gallant, J., & Libben, G. (2019). No lab, no problem: Designing lexical comprehension and production experiments using PsychoPy3. *The Mental Lexicon*, *14*(1), 152–168. https://doi.org/https://doi.org/10.1075/ml.00002.gal

Geethanjali, B., Adalarasu, K., & Jagannath, M. (2018). Music induced emotion and music processing in the brain–a review. *Journal of Clinical & Diagnostic Research*, *12*(1). https://doi.org/10.7860/JCDR/2018/30384.11060

Giraldo, O., Garcia, A., & Corcho, O. (2018). A guideline for reporting experimental protocols in life sciences. *PeerJ*, *6*, e4795. https://doi.org/10.7717/peerj.4795

Godwin, J. (1992). *The harmony of the spheres: The Pythagorean tradition in music*. Simon; Schuster.

Goerlich, K. S., Witteman, J., Aleman, A., & Martens, S. (2011). Hearing feelings: Affective categorization of music and speech in alexithymia, an erp study. *PloS one*, *6*(5), e19501. https://doi.org/10.1371/journal.pone.0019501

Goerlich, K. S., Witteman, J., Schiller, N. O., Van Heuven, V. J., Aleman, A., & Martens, S. (2012). The nature of affective priming in music and speech. *Journal of Cognitive Neuroscience*, *24*(8), 1725–1741. https://doi.org/10.1162/jocn_a_00213

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Goodwin, H., Yiend, J., & Hirsch, C. R. (2017). Generalized anxiety disorder, worry and attention to threat: A systematic review. *Clinical Psychology Review*, *54*, 107–122. https://doi.org/10.1016/j.cpr.2017.03.006

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, *52*, 2283–2286. https://doi.org/10.3758/s13428-020-01395-3

Guang, C., Lefkowitz, E., Dillman-Hasso, N., Brown, V., & Strand, J. (2020). Recall of speech is impaired by subsequent masking noise: A replication of Rabbitt (1968) experiment 2. *Auditory Perception & Cognition*, *3*(3), 158–167. https://doi.org/10.1080/25742442.2021.1896908

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–14. https://doi.org/10.1145/3173574.3174023

Hardie, S. M., & Wright, L. (2014). Differences between left-and right-handers in approach/avoidance motivation: Influence of consistency of handedness measures. *Frontiers in Psychology*, *5*, 134. https://doi.org/10.3389/fpsyg.2014.00134

Harrer, M., Cuijpers, P., A, F. T., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide* (1st). Chapman & Hall/CRC Press. https://doi.org/10.1201/9781003107347

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *Dmetar: Companion R package for the guide 'doing meta-analysis in R'* [R package version 0.0.9000]. http://dmetar.protectlab.org/

Harrison, P. M. C., & Jacoby, N. (2020). Psynet: The online human behavior lab of the future. https://www.aesthetics.mpg.de

Harrison, P. M. C., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. *arXiv preprint arXiv:2008.02595.*

Harrison, P. M. C., & Müllensiefen, D. (2018). Development and validation of the computerised adaptive beat alignment test (ca-bat). *Scientific Reports*, *8*(1), 1–19. https://doi.org/10.1038/s41598-018-30318-8

Harrison, P., & Pearce, M. (2018). An energy-based generative sequence model for testing sensory theories of Western harmony. *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 160–167.

Harrison, P., & Pearce, M. (2020). Simultaneous consonance in music perception and composition. *Psychological Review*, *127*(2), 216–244. https://doi.org/10.1037/rev0000169

Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, *51*(4), 1782–1803. https://doi.org/10.3758/s13428-018-1155-z

Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, *17*(4), 279–296. https://doi.org/https://doi.org/10.2307/1165125

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486. https://doi.org/https://doi.org/10.1037/1082-989X.3.4.486

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*(3), 964–988. https://doi.org/10.3758/s13414-016-1274-5

Helmholtz, H. (1885). On the sensations of tone. 1954. *Trans. AJ Ellis. Introduction by H. Margenau. Republication by Dover Publ. Inc.*

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond weird: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2-3), 111–135. https://doi.org/10.1017/S0140525X10000725

Hermans, D., De Houwer, J., & Eelen, P. (2001). A time course analysis of the affective priming effect. *Cognition & Emotion*, *15*(2), 143–165. https://doi.org/10.1080/02699930125768

Hermans, D., Spruyt, A., De Houwer, J., & Eelen, P. (2003). Affective priming with subliminally presented pictures. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *57*(2), 97–114. https://doi.org/10.1037/h0087416

Hermans, D., Vansteenwegen, D., Crombez, G., Baeyens, F., & Eelen, P. (2002a). Expectancy-learning and evaluative learning in human classical conditioning: Affective priming as an indirect and unobtrusive measure of conditioned stimulus valence. *Behaviour Research and Therapy*, *40*(3), 217–234. https://doi.org/10.1016/S0005-7967(01)00006-7

Hermans, D., Vansteenwegen, D., Crombez, G., Baeyens, F., & Eelen, P. (2002b). Expectancy-learning and evaluative learning in human classical conditioning: Affective priming as an indirect and unobtrusive measure of conditioned stimulus valence. *Behaviour Research and Therapy*, *40*(3), 217–234. https://doi.org/https://doi.org/10.1016/S0005-7967(01)00006-7

Herring, D. R., White, K. R., Jabeen, L. N., Hinojos, M., Terrazas, G., Reyes, S. M., Taylor, J. H., & Crites Jr, S. L. (2013). On the automatic activation of attitudes: A quarter century of evaluative priming research. *Psychological Bulletin*, *139*(5), 1062. https://doi.org/10.1037/a0031309

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/https://doi.org/10.1002/sim.1186

Hilbig, B. E. (2016). Reaction time effects in lab-versus web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hindemith, P. (1942). *The Craft of Musical Composition, Vol. 1.* New York: Belwin-Mills.

Hinojosa, J. A., Carretié, L., Méndez-Bértolo, C., Mķguez, A., & Pozo, M. A. (2009). Arousal contributions to affective priming: Electrophysiological correlates. *Emotion*, *9*(2), 164. https://doi.org/10.1037/a0014680

Hitlin, P. (2016). Research in the crowdsourcing age: A case study. https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Hochheimer, C. J., Sabo, R. T., Perera, R. A., Mukhopadhyay, N., & Krist, A. H. (2019). Identifying attrition phases in survey data: Applicability and assessment study. *Journal of Medical Internet Research*, *21*(8), e12811. https://doi.org/10.2196/12811

Holcomb, P. J., & Anderson, J. E. (1993). Cross-modal semantic priming: A time-course analysis using event-related brain potentials. *Language and cognitive processes*, *8*(4), 379–411.

Honing, H. (2021). Lured into listening: Engaging games as an alternative to reward-based crowdsourcing in music research. *Zeitschrift für Psychologie*, *229*(4). https://doi.org/10.1027/2151-2604/a000474

Houwer, J. D., Hermans, D., Rothermund, K., & Wentura, D. (2002). Affective priming of semantic categorisation responses. *Cognition & Emotion*, *16*(5), 643–666. https://doi.org/https://doi.org/10.1080/02699930143000419

Howe, P. D. L., & Lee, S. B. W. (2021). Attribute amnesia in the auditory domain. *Perception*, 03010066211022175. https://doi.org/10.1177/03010066211022175

Hunter, J. P., Saratzis, A., Sutton, A. J., Boucher, R. H., Sayers, R. D., & Bown, M. J. (2014). In meta-analyses of proportion studies, funnel plots were found to be an inaccurate method of assessing publication bias. *Journal of Clinical*

Epidemiology, *67*(8), 897–903. https://doi.org/https://doi.org/10.1016/j.jclinepi.2014.03.003

Huron, D. (2008). *Sweet anticipation: Music and the psychology of expectation*. MIT Press. https://doi.org/10.7551/mitpress/6575.001.0001

Hutchinson, W., & Knopoff, L. (1978). The acoustic component of Western consonance. *Interface*, *7*(1), 1–29. https://doi.org/10.1080/09298217808570246

Hutto, D. D., Robertson, I., & Kirchhoff, M. D. (2018). A new, better bet: Rescuing and revising basic emotion theory. *Frontiers in psychology*, *9*, 360717.

Huziwara, E. M., Cedro, Į. M., Rodrigues, R. M., de Oliveira, T. P., Bortoloti, R., & Jaeger, A. (2023). Affective priming caused by musical chords on human facial expressions. *Psychology of Music*, *51*(2), 541–552. https://doi.org/10.1177/03057356221097996

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception: An Interdisciplinary Journal*, *23*(4), 319–330. https://doi.org/10.1525/mp.2006.23.4.319

Ipsen, C., Kurth, N., & Hall, J. (2021). Evaluating MTurk as a recruitment tool for rural people with disabilities. *Disability and Health Journal*, *14*(1), 100991. https://doi.org/https://doi.org/10.1016/j.dhjo.2020.100991

Izard, C. E., Libero, D. Z., Putnam, P., & Haynes, O. M. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology*, *64*(5), 847. https://doi.org/10.1037/0022-3514.64.5.847

Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370. https://doi.org/10.1016/j.cub.2016.12.031

Jakubowski, K., Belfi, A. M., & Eerola, T. (2021). Phenomenological differences in music-and television-evoked autobiographical memories. *Music Perception: An Interdisciplinary Journal*, *38*(5), 435–455. https://doi.org/10.1525/mp.2021.38.5.435

James, E., Gaskell, M. G., Pearce, R., Korell, C., Dean, C., & Henderson, L. (2020). The role of prior lexical knowledge in children's and adults' word learning from stories. https://doi.org/10.31234/osf.io/vm5ad

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

Ji, L.-J., & Yap, S. (2016). Culture and cognition. *Current Opinion in Psychology*, *8*, 105–111. https://doi.org/10.1002/9781119170174.epcn314

Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On musical dissonance. *Music Perception: An Interdisciplinary Journal*, *30*(1), 19–35. https://doi.org/10.1525/mp.2012.30.1.19

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Citeseer.

Kalyanshetti, S. B., & Vastrad, B. (2013). Effect of handedness on visual, auditory and cutaneous reaction times in normal subjects. *Al Ameen Journal of Medical Sciences*, *6*(3), 278–280.

Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Computers in Human Behavior*, *83*, 243–253. https://doi.org/10.1016/j.chb.2018.02.005

Kanber, E., Lavan, N., & McGettigan, C. (2020, July). Highly accurate and robust identity perception from personally familiar voices. https://doi.org/10.31234/osf.io/hc7q4

Katsuki, F., & Constantinidis, C. (2014). Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, *20*(5), 509–521. https://doi.org/10.1177/1073858413514136

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644. https://doi.org/10.1016/j.neuron.2018.03.044

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PloS One*, *14*(9), e0221802. https://doi.org/10.1371/journal.pone.0221802

Klauer, K. C. (1997). Affective priming. *European Review of Social Psychology*, *8*(1), 67–103. https://doi.org/10.1080/14792779643000083

Klauer, K. C., & Musch, J. (2003). Affective priming: Findings and theories. In *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 19–60). Psychology Press. https://doi.org/10.4324/9781410606853-7

Klinger, M. R., Burton, P. C., & Pitts, G. S. (2000). Mechanisms of unconscious priming: I. response competition, not spreading activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 441. https://doi.org/10.1037/0278-7393.26.2.441

Knight, S., Lavan, N., Torre, I., & McGettigan, C. (2021). The influence of perceived vocal traits on trusting behaviours in an economic game. *Quarterly Journal of Experimental Psychology*, 17470218211010144. https://doi.org/10.1177/17470218211010144

Koelsch, S., Skouras, S., & Lohmann, G. (2018). The auditory cortex hosts network nodes influential for emotion processing: An fMRI study on music-evoked fear and joy. *PloS One*, *13*(1), e0190057. https://doi.org/10.1371/journal.pone.0190057

Koutseff, A., Reby, D., Martin, O., Levrero, F., Patural, H., & Mathevon, N. (2018). The acoustic space of pain: Cries as indicators of distress recovering dynamics in pre-verbal infants. *Bioacoustics*, *27*(4), 313–325. https://doi.org/10.1080/09524622.2017.1344931

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, *72*(5), 341–372. https://doi.org/10.1016/j.pneurobio.2004.03.006

Krumhansl, C. L., & Castellano, M. A. (1983). Dynamic processes in music perception. *Memory & Cognition*, *11*(4), 325–334. https://doi.org/10.3758/BF03202445

Kunst, J. (1942). *Music in Flores: A study of the vocal and instrumental music among the tribes living in Flores*. E.J. Brill.

Lahdelma, I., Armitage, J., & Eerola, T. (2020). Affective priming with musical chords is influenced by pitch numerosity. *Musicae Scientiae*. https://doi.org/https://doi.org/10.1177/1029864920911127

Lahdelma, I., Athanasopoulos, G., & Eerola, T. (2021). Sweetness is in the ear of the beholder: Chord preference across United Kingdom and Pakistani listeners. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.14655

Lahdelma, I., Armitage, J., & Eerola, T. (2022). Affective priming with musical chords is influenced by pitch numerosity. *Musicae Scientiae*, *26*(1), 208–217. https://doi.org/10.1177/1029864920911127

Lahdelma, I., & Eerola, T. (2016). Single chords convey distinct emotional qualities to both naı̧ve and expert listeners. *Psychology of Music*, *44*(1), 37–54. https://doi.org/https://doi.org/10.1177/0305735614552006

Lahdelma, I., & Eerola, T. (2020). Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-65615-8

Lahdelma, I., Eerola, T., & Armitage, J. (2022). Is harmonicity a misnomer for cultural familiarity in consonance preferences? *Frontiers in Psychology, 13: 802385. doi: 10.3389/fpsyg.* https://doi.org/10.3389/fpsyg.2022.802385

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (jatos): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lavan, N., Burston, L. F., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*(3), 576–593. https://doi.org/10.1111/bjop.12348

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners [PMID: 30808271]. *Quarterly Journal of Experimental Psychology*, *72*(9), 2240–2248. https://doi.org/10.1177/1747021819836890

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*, *193*, 104026. https://doi.org/10.1121/1.5137589

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 2404. https://doi.org/10.1038/s41467-019-10295-w

Lavan, N., Kreitewolf, J., Obleser, J., & McGettigan, C. (2021). Familiarity and task context shape the use of acoustic information in voice identity perception. *Cognition*, *215*, 104780. https://doi.org/https://doi.org/10.1016/j.cognition.2021.104780

Lavan, N., Merriman, S. E., Ladwa, P., Burston, L. F., Knight, S., & McGettigan, C. (2020). 'Please sort these voice recordings into 2 identities': Effects of task instructions on performance in voice sorting studies. *British Journal of Psychology*, *111*(3), 556–569. https://doi.org/https://doi.org/10.1111/bjop.12416

Lavan, N., Mileva, M., Burton, A. M., Young, A. W., & McGettigan, C. (2021). Trait evaluations of faces and voices: Comparing within-and between-person variability. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001019

Lavan, N., Mileva, M., & McGettigan, C. (2021). How does familiarity with a voice affect trait judgements? *British Journal of Psychology*, *112*(1), 282–300. https://doi.org/10.1111/bjop.12454

LeDuke, D. O., Borio, M., Miranda, R., & Tye, K. M. (2023). Anxiety and depression: A top-down, bottom-up model of circuit function. *Annals of the New York Academy of Sciences.*

Legget, K. T., Cornier, M.-A., Erpelding, C., Lawful, B. P., Bear, J. J., Kronberg, E., & Tregellas, J. R. (2022). An implicit priming intervention alters brain and behavioral responses to high-calorie foods: A randomized controlled study. *The American Journal of Clinical Nutrition*, *115*(4), 1194–1204. https://doi.org/10.1093/ajcn/nqac009

Lemmens, P. M. C., De Haan, A., Van Galen, G. P., & Meulenbroek, R. G. J. (2007). Emotionally charged earcons reveal affective congruency effects. *Ergonomics*, *50*(12), 2017–25. https://doi.org/10.1080/00140130701524155

Lense, M. D., Gordon, R. L., Key, A. P., & Dykens, E. M. (2013). Neural correlates of cross-modal affective priming by music in Williams syndrome. *Social Cognitive and Affective Neuroscience*, *9*(4), 529–537.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *Sage Open*, *6*(1), 2158244016636433.

Li, M. (2021). The synergistic effects of solutions journalism and corporate social responsibility advertising. *Digital Journalism*, *9*(3), 336–363. https://doi.org/10.1080/21670811.2020.1840407

Li, W., Zinbarg, R. E., & Paller, K. A. (2007). Trait anxiety modulates supraliminal and subliminal threat: Brain potential evidence for early and late processing influences. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(1), 25–36. https://doi.org/10.3758/CABN.7.1.25

Linnavalli, T., Ojala, J., Haveri, L., Putkinen, V., Kostilainen, K., Seppänen, S., & Tervaniemi, M. (2020). Musical expertise facilitates dissonance detection on behavioral, not on early sensory level. *Music Perception: An Interdisciplinary Journal*, *38*(1), 78–98. https://doi.org/10.1525/mp.2020.38.1.78

Liu, X., Xu, Y., Alter, K., & Tuomainen, J. (2018). Emotional connotations of musical instrument timbre in comparison with emotional speech prosody: Evidence from acoustics and event-related potentials. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00737

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, *82*(1), 5. https://doi.org/10.1037/0022-3514.82.1.5

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

MacInnis, C. C., Boss, H. C., & Bourdage, J. S. (2020). More evidence of participant misrepresentation on MTurk and investigating who misrepresents. *Personality and Individual Differences*, *152*, 109603. https://doi.org/10.1016/j.paid.2019.109603

Maher, T. F. (1976). "Need for resolution" ratings for harmonic musical intervals: A comparison between Indians and Canadians. *Journal of Cross-Cultural Psychology*, *7*(3), 259–276. https://doi.org/https://doi.org/10.1177/002202217673001

Maher, T. F., & Jairazbhoy, N. A. (1975). Need for resolution of musical intervals: Part 1. Static context. *Sangeet Natak Akademi*, *36*, 5–20.

Maier, M. A., Berner, M. P., & Pekrun, R. (2003). Directonality of affective priming: Effects of trait anxiety and activation level. *Experimental Psychology*, *50*(2), 116. https://doi.org/https://doi.org/10.1026//1618-3169.50.2.116

Mallik, A., & Russo, F. A. (2022). The effects of music & auditory beat stimulation on anxiety: A randomized clinical trial. *PloS ONE*, *17*(3), e0259312. https://doi.org/https://doi.org/10.1371/journal.pone.0259312

Marin, M. M., Gingras, B., & Bhattacharya, J. (2012). Crossmodal transfer of arousal, but not pleasantness, from the musical to the visual domain. *Emotion (Washington, D.C.)*, *12*(3), 618–631. https://doi.org/10.1037/a0025020

Marin, M. M., Thompson, W. F., Gingras, B., & Stewart, L. (2015). Affective evaluation of simultaneous tone combinations in congenital amusia. *Neuropsychologia*, *78*, 207–220. https://doi.org/10.1016/j.neuropsychologia.2015.10.004

Mattes, J., & Cantor, J. (1982). Enhancing responses to television advertisements via the transfer of residual arousal from prior programming. *Journal of Broadcasting & Electronic Media*, *26*(2), 553–566. https://doi.org/10.1080/08838158209364024

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual differences reveal the basis of consonance. *Current Biology*, *20*(11), 1035–1041. https://doi.org/10.1121/1.3384926

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, *535*(7613), 547–550. https://doi.org/https://doi.org/10.1038/nature18635

McPherson, M. J., Dolan, S. E., Durango, A., Ossandon, T., Valdés, J., Undurraga, E. A., Jacoby, N., Godoy, R. A., & McDermott, J. H. (2020). Perceptual fusion of musical notes by native amazonians suggests universal representations of musical intervals. *Nature Communications*, *11*(1), 2786. https://doi.org/10.1038/s41467-020-16448-6

Mehr, S. A., Krasnow, M. M., Bryant, G. A., & Hagen, E. H. (2021). Origins of music in credible signaling. *Behavioral and Brain Sciences*, *44*, e60. https://doi.org/10.1017/S0140525X20000345

Meier, B. P., Sellbom, M., & Wygant, D. B. (2007). Failing to take the moral high ground: Psychopathy and the vertical representation of morality. *Personality and Individual Differences*, *43*(4), 757–767. https://doi.org/10.1016/j.paid.2007.02.001

Mellis, A. M., & Bickel, W. K. (2020). Mechanical turk data collection in addiction research: Utility, concerns and best practices. *Addiction*, *115*(10), 1960–1968. https://doi.org/10.1111/add.15032

Messner, G. F. (1981). The two-part vocal style on Baluan Island Manus province, Papua New Guinea. *Ethnomusicology*, *25*(3), 433–446. https://doi.org/https://doi.org/10.2307/851553

Meyer, C., Padmala, S., & Pessoa, L. (2019). Dynamic threat processing. *Journal of Cognitive Neuroscience*, *31*(4), 522–542. https://doi.org/10.1162/jocn_a_01363

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227. https://doi.org/10.1037/h0031564

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, 1–12. https://doi.org/10.3758/s13428-020-01514-0

Miyazaki, K. (1989). Absolute pitch identification: Effects of timbre and pitch region. *Music Perception: An Interdisciplinary Journal*, *7*(1), 1–14. https://doi.org/10.2307/40285445

Mogg, K., Bradley, B., Miles, F., & Dixon, R. (2004). Brief report time course of attentional bias for threat scenes: Testing the vigilance-avoidance hypothesis. *Cognition and emotion*, *18*(5), 689–700.

Mogg, K., & Bradley, B. P. (1999). Selective attention and anxiety: A cognitive-motivational perspective. *Handbook of cognition and emotion*, 145–170.

Moore, B. C., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, *74*(3), 750–753. https://doi.org/10.1121/1.389861

Moro, S. S., & Steeves, J. K. (2021). Lack of affective priming indicates attitude-behaviour discrepancy for covid-19 affiliated words. *Scientific Reports*, *11*(1), 21912. https://doi.org/10.1038/s41598-021-01210-9

Moro, S. S., & Steeves, J. K. (2023). Assessment of implicit covid-19 attitudes using affective priming for pro-vaccine and vaccine-hesitant individuals. *Journal of*

Health Psychology, 13591053231176261. https://doi.org/10.1177/1359105323117 6261

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105. https://doi.org/https://doi.org/10.1037/1082-989X.7.1.105

Moss, A., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it [accessed February 2019]. https://t.ly/G9YKZ

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138. https://doi.org/10.1017/XPS.2015.19

Nacke, L. E., & Deterding, C. S. (2017). The maturing of gamification research. *Computers in Human Behaviour*, *71*, 450–454. https://doi.org/doi.org/10.1016/j .chb.2016.11.062

Nakien, A. (2003). On instrumental origins of Lithuanian polymodal "sutartins". *Studia Musicologica*, 159–168.

Neisser, U. (2014). *Cognitive psychology: Classic edition*. Psychology Press.

Niiniluoto, I. (2011). The development of the Hintikka Program. In *Handbook of the history of logic* (pp. 311–356, Vol. 10). Elsevier. https://doi.org/10.1016/B978-0 -444-52936-7.50009-4

Nilsson, U. (2008). The anxiety-and pain-reducing effects of music interventions: A systematic review. *AORN Journal*, *87*(4), 780–807. https://doi.org/https://doi .org/10.1016/j.aorn.2007.09.013

Njie, Lavan, N., & McGettigan, C. (2021, January). Talker and accent familiarity yield advantages for voice identity perception: A voice sorting study. https://doi .org/10.31234/osf.io/b6ftg

Nordmark, J., & Fahlén, L. E. (1988). Beat theories of musical consonance. *KTH Royal Institute of Technology STL-QPSR*, *29*, 111–122.

O'Brien, A. M., & Schmidt, J. L. (2020). Typical listeners are unable to detect sound quality differences between luxury and value headphones. *Cognition, Brain, Behavior*, *24*(1), 57–74. https://doi.org/https://doi.org/10.1121/at.2022.18.1.58

Öhman, A. (2013). As fast as the blink of an eye: Evolutionary preparedness for preattentive processing of threat. In *Attention and orienting* (pp. 165–184). Psychology Press.

Okon-Singer, H. (2018). The role of attention bias to threat in anxiety: Mechanisms, modulators and open questions. *Current Opinion in Behavioral Sciences*, *19*, 26–30. https://doi.org/10.1016/j.cobeha.2017.09.008

Olive, S., Khonsaripour, O., & Welti, T. A survey and analysis of consumer and professional headphones based on their objective and subjective performances. In: In *Audio Engineering Society Convention 145*. Audio Engineering Society. 2018. http://www.aes.org/e-lib/browse.cfm?elib=19774

Ollen, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings* [Doctoral dissertation, The Ohio State University].

Onnis, R., Dadds, M. R., & Bryant, R. A. (2011). Is there a mutual relationship between opposite attentional biases underlying anxiety? *Emotion*, *11*(3), 582. https://doi.org/10.1037/a0022019

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional manipulation checks: A longitudinal analysis with implications for mturk. *International Journal of Research in Marketing*, *35*(2), 258–269. https://doi.org/https://doi.org/10.1016/j.ijresmar.2018.01.003

Pacheco-Unguetti, A. P., Acosta, A., Callejas, A., & Lupiįńez, J. (2010). Attention and anxiety: Different attentional functioning under state and trait anxiety. *Psychological Science*, *21*(2), 298–304. https://doi.org/10.1177/0956797609359624

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Palan, S., & Schitter, C. (2018). Prolific.ac– A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pallesen, K. J., Brattico, E., Bailey, C., Korvenoja, A., Koivisto, J., Gjedde, A., & Carlson, S. (2005). Emotion processing of major, minor, and dissonant chords: A functional magnetic resonance imaging study. *Annals of the New York Academy of Sciences*, *1060*(1), 450–453. https://doi.org/10.1196/annals.1360.047

Pankovski, T. (2021). Screening for dichotic acoustic context and headphones in online crowdsourced hearing studies. *Canadian Acoustics*, *49*(2). https://jcaa.caa-aca.ca/index.php/jcaa/article/view/3403

Parncutt, R. (1989). *Harmony: A Psychoacoustical Approach*. Springer.

Parncutt, R., & Hair, G. (2011). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, *5*(2), 119–166.

Parncutt, R., Reisinger, D., Fuchs, A., & Kaiser, F. (2019). Consonance and prevalence of sonorities in western polyphony: Roughness, harmonicity, familiarity, evenness, diatonicity. *Journal of New Music Research*, *48*(1), 1–20. https://doi.org/10.1080/09298215.2018.1477804

Partch, H. (1974). *Genesis of a music*. New York: Da Capo Press.

Patterson, R. D. (1976). Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, *59*(3), 640–654. https://doi.org/10.1121/1.380914

Payne, B., Lavan, N., Knight, S., & McGettigan, C. (2020). Perceptual prioritization of self-associated voices. *British Journal of Psychology*, *112*(3), 585–610. https://doi.org/10.1111/bjop.12479

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

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1. https://doi.org/https://doi.org/10.3758/s13428-021-01694-3

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peirce, J. W. (2007). Psychopy—Psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychological Research*, *66*(3), 180–194. https://doi.org/10.1007/s00426-002-0086-5

Persichetti, V. (1961). *Twentieth century harmony*. WW Norton New York.

Pfordresher, P. Q., & Demorest, S. M. (2020). Construction and validation of the Seattle Singing Accuracy Protocol (ssap): An automated online measure of singing accuracy. In *The Routledge Companion to Interdisciplinary Studies in Singing* (pp. 322–333). Routledge. https://doi.org/https://doi.org/10.4324/9781315163734-24

Pfordresher, P. Q., & Demorest, S. M. (2021). The prevalence and correlates of accurate singing. *Journal of Research in Music Education*, *69*(1), 5–23. https://doi.org/10.1177/0022429420951630

Pinto, Y., van der Leij, A. R., Sligte, I. G., Lamme, V. A., & Scholte, H. S. (2013). Bottom-up and top-down attention are independent. *Journal of Vision*, *13*(3), 16–16. https://doi.org/10.1167/13.3.16

Pollet, T. V., & Saxton, T. K. (2019). How diverse are the samples used in the journals 'Evolution & Human Behavior'and 'Evolutionary Psychology'? *Evolutionary Psychological Science*, *5*(3), 357–368.

Posner, M. I. (1978). *Chronometric explorations of mind.* Lawrence Erlbaum.

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2020). Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences*, *117*(21), 11364–11367. https://doi.org/10.1073/pnas.2004163117

Prince, J. B., Thompson, W. F., & Schmuckler, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1598. https://doi.org/10.1037/a0016456

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*(5), 410–430. https://doi.org/10.1002/bs.3830120511

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rainait-Vyinien, D. (2012). The revival of Lithuanian polyphonic Sutartins songs in the late 20th and early 21st century. *Res Musica*, *4*, 97–115.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Ragozzine, F. (2011). Cross-modal affective priming with musical stimuli: Effect of major and minor triads on word-valence categorization. *Journal of ITC Sangeet Research Academy*, *25*, 8–24.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020). More time for science: Using testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. *Progressive Brain Research*, *253*, 243–262. https://doi.org/10.1016/bs.pbr.2020.06.005

Robinson, M. D., Ode, S., Moeller, S. K., & Goetz, P. W. (2007). Neuroticism and affective priming: Evidence for a neuroticism-linked negative schema. *Personality and individual differences*, *42*(7), 1221–1231. https://doi.org/10.1016/j.paid.2006.09.027

Robinson, M. D., & Tamir, M. (2005). Neuroticism as mental noise: A relation between neuroticism and reaction time standard deviations. *Journal of Personality and Social Psychology*, *89*(1), 107–114. https://doi.org/10.1037/0022-3514.89.1.107

Rodd, J. (2019). How to maintain data quality when you can't see your participants. *Observer (Association for Psychological Science)*. https://www.psychologicalscience.org/observer/how-to-maintain-data-quality-when-you-cant-see-your-participants

Roederer, J. G. (2008). *The physics and psychophysics of music: An introduction*. Springer Science & Business Media. https://doi.org/10.1119/1.1987821

Rohr, M., & Wentura, D. (2022). How emotion relates to language and cognition, seen through the lens of evaluative priming paradigms. *Frontiers in Psychology*, *13*, 911068.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161. https://doi.org/10.1037/h0077714

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., & Milland, K. (2015). We are Dynamo: Overcoming stalling and friction in collective action for crowd workers. *Proceedings of the 33rd annual ACM Conference on Human Factors in Computing Systems*, 1621–1630. https://doi.org/10.1145/2702123.2702508

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4). https://doi.org/10.3390/brainsci10040251

Scerrati, E., D'Ascenzo, S., Nicoletti, R., Villani, C., & Lugli, L. (2022). Assessing interpersonal proximity evaluation in the covid-19 era: Evidence from the affective priming task. *Frontiers in Psychology*, *13*, 901730. https://doi.org/10.3389/fpsyg.2022.901730

Scharf, B. (1970). Critical bands. In J. Tobias (Ed.), *Foundations of modern auditory theory* (pp. 157–202, Vol. 1). Academic.

Scherer, L. D., & Lambert, A. J. (2009). Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of Personality and Social Psychology*, *97*(3), 383. https://doi.org/10.1037/a0015844

Scherer, L. D., & Larsen, R. J. (2011). Cross-modal evaluative priming: Emotional sounds influence the processing of emotion words. *Emotion*, *11*(1), 203–208. https://doi.org/10.1037/a0022588

Schmidtke, D., Gagné, C. L., Kuperman, V., Spalding, T. L., & Tucker, B. V. (2018). Conceptual relations compete during auditory and visual compound word recognition. *Language, Cognition and Neuroscience*, *33*(7), 923–942. https://doi.org/10.1080/23273798.2018.1437192

Schwartz, J. W., Engelberg, J. W., & Gouzoules, H. (2019). Was that a scream? Listener agreement and major distinguishing acoustic features. *Journal of Nonverbal Behavior*, 1–20. https://doi.org/10.1007/s10919-019-00325-y

Semuels, A. (2018). The internet is enabling a new kind of poorly paid hell. *The Atlantic*. https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

Sheehan, K. B. (2018). Crowdsourcing research: Data collection with Amazon's Mechanical Turk. *Communication Monographs*, *85*(1), 140–156. https://doi.org/10.1080/03637751.2017.1342043

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, *48*(2), 553–566. https://doi.org/10.3758/s13428-015-0599-7

Smith, J. O., & Abel, J. S. (1999). Bark and ERB bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, *7*(6), 697–708. https://doi.org/10.1109/89.799695

Smith, N. A., Sabat, I. E., Martinez, L. R., Weaver, K., & Xu, S. (2015). A convenient solution: Using MTurk to sample from hard-to-reach populations. *Industrial and Organizational Psychology*, *8*(2), 220. https://doi.org/10.1017/iop.2015.29

Sollberger, B., Rebe, R., & Eckstein, D. (2003). Musical chords as affective priming context in a word-evaluation task. *Music Perception*, *20*(3), 263–282. https://doi.org/10.1525/mp.2003.20.3.263

Soyal, M., Ēelik, N. M., & Pekel, A. (2017). An investigation of the relationship between the reaction times and the state and trait anxiety levels of the athletes. *European Journal of Physical Education and Sport Science*.

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A comparative study of collaborative vs. traditional musical mood annotation. *Proceedings of the 11th Society of Music Information Retrieval Conference (ISMIR)*, *104*, 549–554.

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language*, *61*(1), 37–54. https://doi.org/10.1016/j.jml.2009.03.004

Spruyt, A., Houwer, J. D., Everaert, T., & Hermans, D. (2012). Unconscious semantic activation depends on feature-specific attention allocation. *Cognition*, *122*(1), 91–95. https://doi.org/https://doi.org/10.1016/j.cognition.2011.08.017

Spruyt, A., & Tibboel, H. (2015). On the automaticity of the evaluative priming effect in the valent/non-valent categorization task. *PloS one*, *10*(3), e0121564. https://doi.org/https://doi.org/10.1371/journal.pone.0121564

Spruyt, A., Tibboel, H., De Schryver, M., & De Houwer, J. (2017). Automatic stimulus evaluation depends on goal relevance. *Emotion*, *18*(3), 332–341. https://doi.org/10.1037/emo0000361

Stamps, L. E., Fehr, L. A., & Lewis, R. A. (1979). Differential effects of state and trait anxiety on heart rate responses and reaction time. *Biological Psychology*, *8*(4), 265–272. https://doi.org/10.1016/0301-0511(79)90008-5

Steinbeis, N., & Koelsch, S. (2008). Comparing the processing of music and language meaning using eeg and fmri provides evidence for similar and distinct neural representations. *PloS One*, *3*(5), e2226. https://doi.org/10.1371/journal.pone.0002226

Steinbeis, N., & Koelsch, S. (2011). Affective priming effects of musical sounds on the processing of word meaning. *Journal of cognitive neuroscience*, *23*(3), 604–621. https://doi.org/10.1162/jocn.2009.21383

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Storbeck, J., & Clore, G. L. (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass*, *2*(5), 1824–1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x

Stoycheff, E. (2016). Please participate in part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations*, *9*, 2059799116672879. https://doi.org/10.1177/2059799116672879

Strmiska, M. (2005). The music of the past in modern Baltic Paganism. *Nova Religio*, *8*(3), 39–58.

Stumpf, C. (1898). Konsonanz und dissonanz (consonance and dissonance). *Beiträge zur Akustik und Musikwissenschaft*, *1*, 1–108.

Sussman, T. J., Jin, J., & Mohanty, A. (2016). Top-down and bottom-up factors in threat-related perception and attention in anxiety. *Biological psychology*, *121*, 160–172.

Sussman, T. J., Jin, J., & Mohanty, A. (2020). Chapter 10 - the impact of top-down factors on threat perception biases in health and anxiety. In T. Aue & H. Okon-Singer (Eds.), *Cognitive biases in health and psychiatric disorders* (pp. 215–241). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-816660-4.00010-6

Tekman, H. G., & Bharucha, J. J. (1998). Implicit knowledge versus psychoacoustic similarity in priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 252.

Tenderini, M. S., de Leeuw, E., Eilola, T. M., & Pearce, M. T. (2022). Reduced cross-modal affective priming in the L2 of late bilinguals depends on L2 exposure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(2), 284. https://doi.org/10.1037/xlm0000889

Terhardt, E. (1974). On the perception of periodic sound fluctuations (roughness). *Acta Acustica united with Acustica*, *30*(4), 201–213.

Terhardt, E. (1984). The concept of musical consonance: A link between music and psychoacoustics. *Music Perception: An Interdisciplinary Journal*, *1*(3), 276–295. https://doi.org/10.2307/40285261

Tervaniemi, M., Kruck, S., De Baene, W., Schröger, E., Alter, K., & Friederici, A. D. (2009). Top-down modulation of auditory processing: Effects of sound context, musical expertise and attentional focus. *European Journal of Neuroscience*, *30*(8), 1636–1642. https://doi.org/10.1111/j.1460-9568.2009.06955.x

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216. https://doi.org/https://doi.org/10.1016/S0895-4356(99)00161-4

Tierney, A., Patel, A., Jasmin, K., & Breen, M. (2021). Individual differences in perception of the speech-to-song illusion are linked to musical aptitude but not musical experience. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/axumr

Tiihonen, M., Haumann, N. T., Shtyrov, Y., Vuust, P., Jacobsen, T., & Brattico, E. (2024). The impact of crossmodal predictions on the neural processing of aesthetic stimuli. *Philosophical Transactions of the Royal Society B*, *379*(1895), 20220418. https://doi.org/10.1098/rstb.2022.0418

Tramo, M. J., Cariani, P. A., Delgutte, B., & Braida, L. D. (2001). Neurobiological foundations for the theory of harmony in Western tonal music. *Annals of the New York Academy of Sciences*, *930*(1), 92–116. https://doi.org/10.1111/j.1749-6632.2001.tb05727.x

Treiblmaier, H., Putz, L.-M., & Lowry, P. B. (2018). Setting a definition, context, and theory-based research agenda for the gamification of non-gaming applications. *Association for Information Systems Transactions on Human-Computer Interaction (THCI)*, *10*(3), 129–163. https://doi.org/10.17705/1thci.00107

Trevor, C., Arnal, L. H., & Frühholz, S. (2020). Terrifying film music mimics alarming acoustic feature of human screams. *The Journal of the Acoustical Society of America*, *147*(6), EL540–EL545. https://doi.org/10.1121/10.0001459

Turner, A. M., Kirchhoff, K., & Capurro, D. (2012). Using crowdsourcing technology for testing multilingual public health promotion materials. *Journal of Medical Internet Research*, *14*(3), e79. https://doi.org/10.2196/jmir.2063

Tzavella, L., Maizey, L., Lawrence, A. D., & Chambers, C. D. (2020). The affective priming paradigm as an indirect measure of food attitudes and related choice behaviour. *Psychonomic Bulletin & Review*, *27*, 1397–1415. https://doi.org/10.3758/s13423-020-01764-1

van Breen, J. A., Spears, R., Kuppens, T., & de Lemus, S. (2018). Subliminal gender stereotypes: Who can resist? *Personality and Social Psychology Bulletin*, *44*(12), 1648–1663. https://doi.org/10.1177/0146167218771895

Vassilakis, P. N. (2001). *Perceptual and physical properties of amplitude fluctuation and their musical significance* [Doctoral dissertation, University of California, Los Angeles].

Vassilakis, P., & Kendall, R. A. (2008). Auditory roughness profiles and musical tension/release patterns in a Bosnian ganga song. *The Journal of the Acoustical Society of America*, *124*(4), 2448–2448. https://doi.org/https://doi.org/10.1121/1.4782594

Västfjäll, D. (2001). Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, *5*(1_suppl), 173–211. https://doi.org/10.1177/10298649020050S107

Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in associative and categorical priming: A diffusion model analysis. *Journal of Experimental Psychology: General*, *142*(2), 536–559. https://doi.org/10.1037/a0029459

Vyinien, D. (2002). Lithuanian Schwebungsdiaphonie and its south and east European parallels. *The World of Music*, 55–77.

Wang, Y., Zhang, Z., Bai, L., Lin, C., Osinsky, R., & Hewig, J. (2017). Ingroup/outgroup membership modulates fairness consideration: Neural signatures from ERPs and EEG oscillations. *Scientific Reports*, *7*, 39827. https://doi.org/10.1038/srep39827

Wang, Y., Shen, G., Guo, H., Tang, X., & Hamade, T. (2013). Roughness modelling based on human auditory perception for sound quality evaluation of vehicle interior noise. *Journal of Sound and Vibration*, *332*(16), 3893–3904. https://doi.org/10.1016/j.jsv.2013.02.030

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/https://doi.org/10.3758/s13428-012-0314-x

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158. https://doi.org/10.1037/h0074428

Wentura, D. (1999). Activation and inhibition of affective information: Evidence for negative priming in the evaluation task. *Cognition and Emotion*, *13*(1), 6591. https://doi.org/10.1080/026999399379375

Wentura, D. (2000). Dissociative affective and associative priming effects in the lexical decision task: Yes versus no responses to word targets reveal evaluative judgment tendencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 456. https://doi.org/10.1037/0278-7393.26.2.456

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482. https://doi.org/10.1007/BF03395630

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilkerson, J. M., Iantaffi, A., Grey, J. A., Bockting, W. O., & Rosser, B. S. (2014). Recommendations for internet-based qualitative health research with hard-to-reach populations. *Qualitative Health Research*, *24*(4), 561–574. https://doi.org/10.1177/1049732314524635

Williams, J. M. G., Watts, F. N., MacLeod, C., & Mathews, A. (1988). *Cognitive psychology and emotional disorders.* John Wiley & Sons.

Winkielman, P., Knutson, B., Paulus, M., & Trujillo, J. L. (2007). Affective influence on judgments and decisions: Moving towards core mechanisms. *Review of General Psychology*, *11*(2), 179. https://doi.org/10.1037/1089-2680.11.2.179

Wittenbrink, B., & Schwarz, N. (2007). *Implicit measures of attitudes.* Guilford Publications.

Wolff, D., MacFarlane, A., & Weyde, T. (2015). Comparative music similarity modelling using transfer learning across user groups. *Proceedings of the 15th Society of Music Information Retrieval Conference (ISMIR)*, 24–30.

Woods, K. J., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, *115*(14), E3313–E3322. https://doi.org/10.1073/pnas.1801614115

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Yamada, M., & Decety, J. (2009). Unconscious affective processing and empathy: An investigation of subliminal priming on the detection of painful facial expressions. *Pain*, *143*(1-2), 71–75. https://doi.org/10.1016/j.pain.2009.01.028

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception, & Psychophysics*, *81*(2), 558–570. https://doi.org/https://doi.org/10.3758/s13414-018-1635-3

Yuan, J., Chen, J., Yang, J., Ju, E., Norman, G. J., & Ding, N. (2014). Negative mood state enhances the susceptibility to unpleasant events: Neural correlates from a music-primed emotion classification task. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0089844

Zack, E. S., Kennedy, J., & Long, J. S. (2019). Can nonprobability samples be used for social science research? A cautionary tale. *Survey Research Methods*, *13*(2), 215–227. https://doi.org/10.18148/srm/2019.v13i2.7262

Zhang, J. D., & Schubert, E. (2019). A single item measure for identifying musician and nonmusician categories based on measures of musical sophistication. *Music Perception: An Interdisciplinary Journal*, *36*(5), 457–467. https://doi.org/10.1525/mp.2019.36.5.457

Zhang, Q., Kong, L., & Jiang, Y. (2012). The interaction of arousal and valence in affective priming: Behavioral and electrophysiological evidence. *Brain Research*, *1474*, 60–72. https://doi.org/10.1016/j.brainres.2012.07.023

Zhang, Q., Lawson, A., Guo, C., & Jiang, Y. (2006). Electrophysiological correlates of visual affective priming. *Brain Research Bulletin*, *71*(1), 316–323. https://doi.org/10.1016/j.brainresbull.2006.09.023

Zhao, H., Jiang, J., Zhou, L., & Jiang, C. (2019). Role of the human mirror system in automatic processing of musical emotion: Evidence from EEG. *Acta Psychologica Sinica*, *51*(7), 795. https://doi.org/10.3724/SP.J.1041.2019.00795

Zhou, L., Liu, F., Jiang, J., & Jiang, C. (2019). Impaired emotional processing of chords in congenital amusia: Electrophysiological and behavioral evidence. *Brain and Cognition*, *135*, 103577. https://doi.org/https://doi.org/10.1016/j.bandc.2019.06.001

Zillmann, D. (1983). Transfer of excitation in emotional behavior. *Social psychophysiology: A sourcebook*, 215–240.

Zillmann, D., Katcher, A. H., & Milavsky, B. (1972). Excitation transfer from physical exercise to subsequent aggressive behavior. *Journal of Experimental Social Psychology*, *8*(3), 247–259. https://doi.org/10.1016/S0022-1031(72)80005-2