sempre:
Society for Education, Music and Psychology Research
MUSIC & SCIENCE

Research Article

Music & Science
Volume 8: 1–15

DOI: 10.1177/20592043251341705
journals.sagepub.com/home/mns
Sage

Discrimination in Evaluation: A Call for Greater Attention to Issues of Racial Discrimination in Experimental Musical Performance Evaluation Research

Cayenna Ponchione-Bailey[1]

Abstract

This article provides a summary and critical review of the experimental perceptual studies within music psychology which have investigated the effect of evaluators' perceived race of performers and the interaction with evaluators' race on performance evaluations. With racial inequities continuing to be a major issue within historically-European classical music performance, controlled experimental studies can provide an important source of empirical data on which to base productive interventions. This article identifies a number of issues in the existing research which could benefit from future interdisciplinary collaboration. These include the need to define and control for racialized characteristics of performance variables, the need for further consideration of potential racial bias in evaluative frameworks and concepts, the need to address the racialization of research subjects and participants, and the need to avoid reaffirming harmful stereotypes through study designs and unsupported conjectures.

Keywords

Music perception, musical performance, music psychology, racial bias, racial discrimination

Submission date: 9 July 2021; Acceptance date: 4 April 2025

Introduction

Extensive research within the field of music perception has confirmed that nominally extrinsic information shapes viewers' evaluations and experiences of musical performance. This includes performer gestures and facial expressions (e.g., Bland & Cho, 2021; Davidson, 1993; Morrison et al., 2014; Nápoles et al., 2021; Nusseck & Wanderley, 2009; Vuoskoski et al., 2016), attire, physical attractiveness, and perceived age and gender (male/female) (e.g., Behne & Wöllner, 2011; Griffiths, 2008; Harrington, 2018; Wapnick et al., 1997, 1998), order of performance, familiarity with the music, and contextual information including professional status (Chmurzyńska, 2023; Kroger & Margulis, 2017). However, despite the substantial evidence that racially minoritized musicians are underrepresented and experience racial discrimination and harassment within Anglo-European HECM industries (Brown, 2020a; Cox & Kilshaw [Arts Council England], 2021; Di Laccio [Donne in Music], 2022; Doeser [League of American Orchestras], 2016; Flagg, 2020; Gittens et al. [Black Lives in Music], 2021; Griffiths 2020; Kawabata, 2023; Koh, 2021; Kolbe, 2022; Scharff, 2015; Spence [Music Mark], 2021; Vann, 2018; Williams & Bain [Independent Society of Musicians], 2022; Yoshihara, 2021), beyond a few studies in the 1990s which looked at children's listening preferences and musical identification (Killian, 1990; McCrary, 1993; Morrison, 1998; see Bermingham, 2000 for a review of research up to that date), only a small number of investigations have sought to investigate the effect of evaluators' perceived race of performers and the interaction with evaluators' race on performance evaluations (Cheek, 2007; Clauhs, 2013; Davidson

[1] Music, University of Oxford, Oxford, UK

Corresponding author:
Cayenna Ponchione-Bailey, Music, University of Oxford, Oxford, UK.
Email: cayenna.ponchione@music.ox.ac.uk

& Edgar, 2003; Elliott, 1995; Peynircioğlu et al., 2018; VanWeelden, 2004; VanWeelden & McGee, 2007). This lack of attention to racialization in music performance research is reflected in the field of psychology more generally (Roberts et al., 2020).

In this article I summarize and critically review five peer-reviewed articles reporting on experimental perceptual studies in music psychology that investigated the effect of performers and evaluators race in HECM performance evaluations (Davidson & Edgar, 2003; Elliott, 1995; Peynircioğlu et al., 2018; VanWeelden, 2004; and VanWeelden & McGee, 2007), and two doctoral dissertations (Cheek, 2007; and Clauhs, 2013). Identifying a number of challenges raised by the studies, I argue for the need for further interdisciplinary research in this area to address these issues. These include the need to define and control for racialized characteristics of performance variables, the need for further consideration of potential racial bias in evaluative frameworks and concepts, the need to address the racialization of research subjects and participants, and the need to avoid reaffirming harmful stereotypes through study designs and unsupported conjectures. In doing so, this discussion aims to contribute to the current dialogue about racism in music psychology research and within the field of music psychology (Baker et al., 2020; Eagle et al., 2023; Jacoby et al., 2020; Suavé et al., 2023), and the wider ongoing conversations on decolonization and racism in music studies (e.g., Brown, 2020b; Ewell, 2020; Morrison, 2019; Pistorius, 2019; Tan, 2021).

In line with recommendations for greater transparency about researchers' identities and their positionality regarding their areas of inquiry within the field of psychology (Roberts et al., 2020) and music science (Suavé et al., 2023) more specifically, I offer a brief positionality statement:

> I identify as a cis-female White (European American) early career researcher and professional orchestral conductor living in the United Kingdom. My research specialism is in the social psychology of orchestral performance with a current focus on orchestral cultures in the United Kingdom and of Afghanistan. This review article arises from my preparatory work to undertake collaborative research in perceptual studies which explores the effect of performer and viewer-listener characteristics on the evaluations of orchestral performers and performances.

The topic is sensitive, and I admit to feeling some vulnerability navigating these issues while critically reviewing others' work. However, I feel it is essential as a research community that we take these risks to open up dialogue about such crucial topics, and I welcome challenges and corrections to the analysis offered here. In the following section I outline my rationale for the use of the terms "race" and "racialized", the phrase "historically-European classical music", and the limitations of this review with regard to sex, gender, and international scope.

Notes on Terminology and Survey Limitations

The Terms "Race" and "Racialization"

All the studies reviewed here used the term "race" to describe a demographic variable in their research designs. It is now widely accepted that there is no biological basis for race, but rather that race categories have been created for social, political, and economic purposes often involving the exercise of power of one racialized group over another (Kendi, 2019; Sussman, 2014). "Racialization"—the process by which groups become misunderstood for being biological groups (Hochman, 2019; Martinez, 2023)—is multifaceted, with one dimension involving judgments based on the somatic features of individuals. Indeed, central to the designs of the studies reviewed here is the assumed or explicit racialization of subjects by their visible physical attributes, with particular attention to skin color and facial characteristics. Thus, it is perhaps more accurate to understand this research as studying evaluations of music performances by and of "racialized groups". Martinez (2023) advocates for moving away from the word "race", and instead utilizing the language of racialization in perceptual research, which I adopt in this article. However, the studies reviewed here use the term "race" to describe the variable in their study designs and I use the term when discussing their work, while noting the danger in reifying the concept.

I have also chosen to retain the original race category terminology employed by authors in discussions of their work, with the exception of the capitalized forms of b/Black and White (if not capitalized already) alongside racial terminology that is typically capitalized such as Asian, East Asian, African American, Afro-Cuban, and White European unless contained within a direct quotation. I use the term Black (without b/) when discussing this racialized category in North American contexts. These choices follow the argument that "to not name "White" as a race is…an anti-Black act which frames Whiteness as both neutral and the standard" (Nguyễn & Pendleton, 2020), and acknowledges the North American origins of the capitalization of Black and therefore its inadequacy in representing all b/Black identities (Douglas, 2021). An alternative approach, advocated by Martinez, and one which future authors may wish to adapt, is to not capitalize any race categories in order to resist their reification within the text (Martinez, 2023, p. 190). I acknowledge that racial categories, their terminology, and their appropriateness will differ by geographical location, and that this terminology (and our collective understanding of its meanings) will change over time.

The Basis for the Use of "Historically-European Classical Music" (HECM)

The adoption of this terminology marks a move away from subsuming all people in the Western hemisphere into

European-heritage cultures, and signals respect for the many forms that HECM now takes across the globe —"'western art music' is (solely) western no longer" (Nooshin, 2011, p. 294).

A Note on Sex and Gender

None of the studies reviewed in this article used sex or gender categories other than "male" and "female" in the reporting of participant or subject demographics and some elided between the terms "sex" and "gender" in their writing. In my discussion of the literature, I use gender throughout as it was most frequently employed by all authors. Although it is not the focus of the current article, it is clear that greater attention needs to be paid to the nuances of sex and gender as variables in study designs.

Scope of this Review

This article is focused on HECM practices and music perception research undertaken in the United States and United Kingdom, where the existing research on this topic has been conducted and published in English. It is possible that research and publications in this area exist in other languages, though they have not surfaced in my literature review. It is likely, however, that this is partly indicative of the historical insularity of music psychology research in Europe, America, and Australia (Baker et al., 2020; Jacoby et al., 2020).

In the next section of this article, I provide summaries of the methods and findings of each of the seven studies identified. The level of detail is intended to offer sufficient context to understand the relevance of the findings and provide a snapshot of the methods and scope of the research conducted in this area to date. These summaries are followed by a discussion of issues that arise collectively from these studies, and I conclude with final thoughts on some ways forward for future research.

Summary of Existing Research

Table 1 summarizes the seven studies' subject focus, context, and primary variables, including performers' and evaluators' race and gender, performers' movement and other visual information, and performers' instrument and repertoire. All studies found a significant effect of the parameter of the perceived race of performers, with individual studies also showing significant interactions with the race and gender of evaluators, performers' instruments, repertoire, gender, and gestures. Despite these findings, little research appears to have been done to follow up this work, a situation reflective of the general replication crisis in music psychology more broadly (Frieler et al., 2013). The majority of this research has been conducted in the United States and the United Kingdom—following the historical Euro-centric trends in music psychology research (Jacoby et al., 2020)—with the exception of Peynircioğlu et al., which included research participants in China. The studies have also taken place predominantly in educational contexts and focused on a small number of solo instruments (piano, flute, and trumpet) and choral settings. Each study will now be summarized in turn beginning with solo instrumental performance (Elliott, Davidson and Edgar, Peynircioğlu et al., and Clauhs), followed by choral performance (VanWeelden, VanWeelden and McGee, and Cheek).

Solo Instrumental Performance

Elliott (1995) investigated the effect of secondary school student performers' gender (male/female), race (Black/White), and instrument (flute/trumpet), on the evaluations of the players' performances by undergraduate music education students in the United States. Eight secondary school student musicians (4 trumpet and 4 flute players, one each Black male, White male, Black female, and White female) were video-recorded performing the same piece, Etude #10 from *The Watkins-Farnum Performance Scale, Form A*. To control for performance quality, two advanced music students, one trumpeter and one flautist, were recorded to overdub the student videos. Video recordings were taken at a distance so that players' technique (embouchures, fingers) was not readily discernible, and performer order was partially randomized into four stimulus videotapes, each with the four trumpet players followed by the four flute players. Eighty-eight American undergraduate and graduate-level music education students (54 male, 34 female; 47 White, 34 Black, and 6 Asian) evaluated the performances on a grading scale of 1–9 with the option of giving written comments.

An ANOVA revealed strongly significant main effects for race and instrument, with Black students' performances rated significantly lower than the White students' performances and flautists scoring significantly higher than trumpeters. Significant interactions between gender and race and race and instrument were also found: White females scored lower than White males; Black males scored lower than Black females; Black trumpeters were scored lower than Black flautists; and White trumpeters scored higher than White flautists (Elliott, 1995, p. 53). Effect sizes were not reported. Elliott observed that since the trumpet players were seen first on all video tapes, low scores for these players may have been influenced by an order effect. Other confounding issues may have included differences in overdubbing precision, performers' movement, and other aspects of the subjects' presentation. Elliot acknowledges the need for considerable replication of this research to come to any conclusions. Concerns related to the racialization of subjects, instruments, and musical materials are addressed further on in this article.

Davidson and Edgar (2003) conducted a United Kingdom-based study to investigate the effects of race and gender in the evaluation of conservatoire-level piano performances, utilizing overdubbed video-recorded performances in addition to four other "modes of performance" to control for the

Table 1. Overview of the studies reviewed in this article, their contexts and primary variables.

Studies reviewed				Primary variables							
Study	Subjects	Country	Context	Performer race	Evaluator race	Performer gender	Evaluator gender	Movement	Visual information	Instrument	Musical genre
Elliott (1995)	Solo trumpeters & flautists	United States	Secondary & tertiary education	Black/White	N/A	Male/Female	Male/Female	N/A	N/A	Trumpet/ Flute	N/A
Davidson and Edgar (2003)	Solo pianists	United Kingdom	Tertiary education	Afro-Caribbean/ White European	Afro-Caribbean/ White European	Male/Female	Male/Female	N/A	Sound only/ Movement only/ Full visual	N/A	N/A
Clauhs (2013)	Solo pianists	United States	Tertiary education	Black/White	N/A	Male/Female	Male/Female	N/A	N/A	N/A	Classical/Jazz
Peynircioğlu et al. (2018)	Solo pianists	United States & China	Tertiary education/ online videos	East Asian/Caucasian	East Asian/Caucasian	Male/Female	N/A	Quantity	Sound only/Partial occlusion/Full visual	N/A	N/A
VanWeelden (2004)	Choral conductors	United States	Tertiary education	Black/White	N/A	N/A	Male/Female	N/A	N/A	N/A	N/A
VanWeelden & McGee (2007)	Choral conductors	United States	Tertiary education	Black/White	Black/White	N/A	N/A	N/A	N/A	N/A	Western art music/ music from Black culture
Cheek (2007)	Choirs	United States	Secondary education	Black/White/Mixed Black & White	Black/White	N/A	N/A	N/A	N/A	N/A	N/A

variables of sound, gesture, and the perceived race of the performers. Their study involved 9 conservatoire-level pianists: 4 Afro-Caribbean (2 male, 2 female), 4 White European (2 male, 2 female) and 1 Indian Asian (male). Only the ratings for the Afro-Caribbean and White European performers were used for analysis—the inclusion of the Indian Asian performer was to distract judges from inferring the aims of the study (Davidson & Edgar, 2003, p. 174). Thirty-six judges were recruited, including 18 White European and 18 Afro-Caribbean (18 male, 18 female), with no intersectional breakdown reported. Citing research demonstrating that judgments are usually made within the first 15–30 s of a performance, a nine-bar, 30-s excerpt from Brahms' Intermezzo in A minor (No. 7) from *8 Klavierstück*, Op. 76 was used as the musical stimulus. Each judge viewed one of four partially randomized VHS (video) tapes of the 9 pianists across the 5 modes of performance: (i) normal lighting, original performance/interpretation (bodies in view), (ii) point-light displays of performers, original performance/interpretation (showing only performers' movement), (iii) normal lighting, overdubbed, (iv) point-light, overdubbed, and (v) sound only, original performance/interpretation; for a total of 45 recordings. Davidson and Edgar's study is unique amongst this research in its use of motion capture technologies and point-light representations to control for performers' physical characteristics. Synchrony in the dubbed conditions ((iii) and (iv)) was achieved through a standardized tempo, with pianists determining their own tempi for their own interpretations ((i), (ii), and (v)). A separate, single pianist's recording of the excerpt at this tempo was used for all 18 of the dubbed conditions. Judges were asked to rate each performance on a scale of 1–7 (poor to excellent) with attention to technique and interpretation (Davidson & Edgar, 2003, pp. 174–176).

The write-up of the findings is difficult to parse. In part this is due to contradictory statements, an apparently missing chart of the interaction between Race of Performer and Performance Mode and an emphasis on main effects that does not take into account whether or not the bodies of musicians could be seen by the evaluators. However, I provide here what appear to me to be the three relevant significant interactions reported from an ANOVA of the data (effect sizes for interactions were not reported). First, female performers were given significantly higher average scores in all performance modes except sound only, suggesting that there was something in their physical movement, at least, that influenced judges' evaluations (Davidson & Edgar, 2003, p. 176). Second, there was an interaction between judges' race and performers' race when analyzing only the dubbed conditions together (natural light, with the performer's body in view; and point-light, showing the performer's movement only). Afro-Caribbean judges rated Afro-Caribbean performers significantly more highly than White European performers, and vice versa, a finding that was attributed to an "in-group bias" (Davidson & Edgar, 2003, pp. 178–179). And third, there was an interaction for performers' race and performance mode between the two dubbed conditions, but the graph is missing, and the details of this interaction are not fully reported. A bit more insight into these findings is provided in the authors' further discussions, where they reveal that:

> It is worth noting here that when Race [of performer] X Judge was observed, White subjects received a higher rating in the normal lighting condition. This result may indicate that although White performers are not preferred when only their body movement is visible in point-light display, they are rated higher than Blacks when race and gender information is available, perhaps due to preconceptions about the abilities of the White pianists. (Davidson & Edgar, 2003, p. 178).

It may be that there is an error here and the authors are referring to the Race X Performance Mode analysis rather than Race X Judge, it is difficult to know.[1] Further issues related to the write-up of this study are discussed later in this article.

Peynircioğlu et al. (2018) investigated what the authors refer to as the "Asian bias" in classical music performance —"that, in Western art music, East Asian performers are proficient in technique but do not have the same expressive skills as Caucasian performers" (Peynircioğlu et al., 2018, p. 296). (The use of the race category "Caucasian" is discussed later in this article.) They bring together research on the correlation between performers' gesture and evaluations of expressiveness in performance with sociological literature on cultural norms of expression. Drawing on research that indicates Asian cultures tend to discourage outward displays of individual emotion and value introverted behavior, they hypothesize that Asian musicians may indeed engage in less movement during performances and that this contributes to the illusion that Asian performers are less expressive. They propose that this perception then leads to a "generalization that perpetuates the stereotype" (Peynircioğlu et al., 2018, p. 296–7). Across four experiments, Peynircioğlu et al. sought to determine: first, if the presence of an Asian bias existed beyond music critic circles (i.e., in the "general public") the extent to which it was in evidence in both Asian and American geographical contexts; and second, if it did exist, to investigate the role of musicians' body movement on viewers' evaluations.

For their stimuli they used a combination of audio-only, video-only, and audio/video excerpts from existing YouTube recordings of Beethoven piano sonatas. To control for variation in the equivalent proficiency of the performances, audio-only versions of the YouTube excerpts were rated (by a professional pianist, a non-professional pianist, and a non-musician). Similarly rated performances were paired with respect to race and gender, to ensure there was equal representation in the quality of performances across both variables.

The first experiment tested whether providing information about performers' gender and race via performers'

names and countries would affect listeners' performance evaluations in an audio-only context. Undergraduate students in the United States (19 female, 5 male; 14 Caucasian, 1 Asian, and 9 "other-race descent") were asked to rate the expression and technique from audio-only conditions of eight performers equally distributed between gender (male/female) and race (Asian/Caucasian) represented by two performers from China and one performer each from Japan, Korea, the United States, Germany, England, and France. A repeated-measures ANOVA showed no significant effects of race or gender in the evaluations of expressivity or technique based on the extrinsic information of name and country alone (Peynircioğlu et al., 2018, p. 298).

In experiment two they restored the video to the same YouTube recordings used in experiment one (audio-only), and engaged two sets of participants, 24 Caucasian undergraduate students from the United States (14 female, 10 male) and 24 Asian participants from China (17 female, 7 male), who evaluated the performances on expression and technique. The aim was to see if any bias emerged, whether it would be specific to the participants in the United States. A mixed-design ANOVA revealed a main effect of race: both Caucasian and Asian participants rated the Caucasian performers significantly higher than the Asian performers on both expression and technique, with the difference in expression ratings being significantly greater than the difference in technique ratings. They observed that although Caucasian evaluators gave higher ratings than their Asian counterparts, the pattern of preference for Caucasian performers was the same for both groups. An analysis of the difference in ratings between experiment one (sound only) and experiment two (video) showed that the interaction in expressivity was significant and approached significance for technique (Peynircioğlu et al., 2018, p. 300).

The third experiment explored the effect of performer movement as a factor in evaluations. To quantify the amount of perceived movement, 12 American undergraduate students (10 female, 2 male; 9 Caucasian and 3 other-race descent) viewed 80 randomly ordered video-only YouTube recordings (equal representation of Caucasian and Asian, male and female performers) and rated their movement on a five-point scale from "extremely low" to "extremely high". The Caucasian performers were given significantly higher movement ratings than the Asian performers (Peynircioğlu et al., 2018, p. 301). Noting that knowledge of the performers' race may have influenced participants' ratings, a follow-on study was conducted using videos with a similar camera angle that did not focus on the performer's hands, and utilizing video object tracking technology to occlude the pianist's head. Twelve (different) American undergraduate students (9 female, 3 male; 9 Caucasian, 1 Asian, and 2 other-race descent) rated 80 randomized video clips on the same scale. When ostensibly no information about performers' race was visible, Caucasian performers were still found to exhibit significantly more movement than Asian performers, including those that were in the videos used in the study's first experiment (Peynircioğlu et al., 2018, p. 301).

In a fourth and final experiment, 40 American undergraduate students (31 female, 9 male; 23 Caucasian, 6 Asian, and 11 other-race descent) were asked to rate 16 full-video and audio YouTube videos with equal numbers of high-moving and low-moving Caucasian and Asian, male and female performers curated from stimuli used in the previous experiments for expressivity and technique. A repeated-measures ANOVA revealed a main effect of movement with high-movement performances receiving significantly higher expressivity ratings, and no interactions for gender or race. High-movement performances also received higher ratings in technique, with Asian pianists scoring significantly higher than their Caucasian counterparts. In an analysis of high-movement performances only, however, Asian pianists also received significantly higher scores for expressivity, as did males as a group (Peynircioğlu et al., 2018, pp. 302–303).

The authors conclude that their findings demonstrate the existence of an Asian bias amongst viewers of HECM performance which persists across cultures and beyond the opinions of White music critics alone. They theorize that the higher expressivity ratings given to Asian performers in the high-movement category of the final study may indeed be evidence of the Asian bias, with viewers apparently impressed by a degree of movement from Asian performers that runs counter to their expectations. They also propose that the higher ratings for the technique of Asian performers in the final study confirms the persistence of the Asian bias in that Asian performers are believed to be more technically proficient than their Caucasian counterparts. Although there are difficulties with this study due to the lack of adequate control of variables including the quantity and nature of performer movement, equivalent proficiency of performances, and issues with the cross-cultural element of the study (to be explored further in my discussion), this study represents an important point in music perception research in that it is the first to address the critical issue of discrimination against Asian performers in the historically-European classical music profession.

Clauhs's United States-based doctoral research (2013) explored how race and gender bias affect college music faculty (staff) evaluations of jazz and classical pianists, and the associations music faculty may make between a performer's gender or race and styles of music (Clauhs, 2013, p. 29). The study's design was built on a principle similar to that of the Implicit Association Test (IAT), which aims to measure implicit bias by participants' response latency in pairing (congruent or incongruent) visual and textual material.[2] In this case, Clauhs sought to use the IAT's concept of timed reactions to obtain "pre-reflective" responses, measuring the correct identification of a short audio excerpt's genre (jazz or classical), rather than participants' response latency. Three hundred and

fifteen participants—applied music faculty (61% male, 90.9% Caucasian) who worked in higher education music departments across the United States—were asked to make two types of judgments: first, to identify the genre of 10 1500-ms video-recorded excerpts (5 jazz, 5 classical); and second, to predict the exam scores for the pianist for each of 10 10-s video-recorded excerpts of the same compositions. Participants were given 15 s to identify the genre after viewing 1500-ms excerpts and 25 s to predict the exam score following the 10-s excerpts. Pilot studies were used to ensure that the musical style of the excerpts was neither 100% obvious nor suggestive of the wrong style all together. The excerpts were selected from pieces listed on the required audition repertoire at the Juilliard School of Music and the Eastman School of Music.

Four computer-randomized treatments of 20 excerpts were curated using the same audio recorded by a classically trained graduate student pianist overdubbed onto video-recorded performances of four classically trained undergraduate pianists (one each Black male, White male, Black female, and White female). The camera was positioned to capture the side of the pianist's face, but hands and keyboard could not be seen. The pianists prepared their performances from sheet music including intentional mistakes written into the parts (the reason for this is not explicated) and were given the target recording to match. Video-editing software was used to align the performer's gesture with the overdubbed audio (e.g., stretching video where necessary). Clauhs employed a between-subjects design, where each participant viewed a single treatment to deter participants suspecting that the study was about the perception of performers' race.

To analyze the results of the genre identification of the 1500 ms excerpts, Clauhs calculated "stereotype scores" for each participant by subtracting the percentage of correctly identified classical excerpts from the percentage of correctly identified jazz excerpts, with lower scores suggesting a stronger association with classical music and higher scores a stronger association with jazz music (Clauhs 2013, p. 50). The study found significant effects of gender and race of the performers with regard to genre stereotypes. An ANOVA revealed a main effect of gender with males significantly more strongly associated with jazz than females (Clauhs, 2013, p. 55). Clauhs also reports significant interactions between gender and race were found using a post-hoc pairwise comparison of stereotype scores, with White and Black females more strongly associated with classical music than Black males, and conversely Black males more strongly associated with jazz than either Black or White females, but not White males (Clauhs, 2013, p. 57–58).

Clauhs (2013) also found significant interactions for gender and race of performers in the predicted exam scores and an interaction with the gender of the evaluators. An ANOVA and post-hoc comparisons showed that Black males scored lower than White males on predicted exam marks in both the jazz and classical conditions, and lower than Black females in the jazz condition (Clauhs, 2013, p. 71, 79). Through ANOVAs and post-hoc comparisons for the classical and jazz conditions, significant interactions between performers gender and race, and evaluator gender were found with male faculty evaluators rating Black males lower than White males and White females in the classical condition, and giving them the lowest scores overall in the jazz condition: lower than White males, White females, and Black females (Clauhs, 2013, pp. 72, 82). Effect sizes were not reported. There were no significant differences in the predicted exam ratings of any gender and race categories when female evaluators' responses were analyzed alone (Clauhs, 2013, p. 90). While an analysis of the effect of evaluators' race would have been desirable, Clauhs explains that it was not possible to analyze the data for the effect of the race of the evaluator because 91% of the respondents from the random sample were White (Clauhs, 2013, p. 58). In summary, Clauhs found that while Black males were more strongly associated with jazz than both Black and White females, they were scored lower than Black and White females and White males in predicted exam marks in the jazz condition.

Choral Performance

Cheek's (2007) doctoral research in the United States investigated the effect of the perceived race of youth choir members (Black/White) and the racial identities and racial attitudes of evaluators in choral competition adjudications. In a between-subjects design, evaluators listened to the same recording of a choral performance while viewing one of three photographs: a homogeneous Black, homogeneous White, or heterogeneous Black and White youth choir, ostensibly representing the choir on the recording.

To create the three photographs, 40 volunteers from local high schools were recruited, asked to self-identify their racial category, and arranged into groups of 19 "singers" with identical choir robes. After a validation survey, digital alterations were made to skin and hair color of two individuals and the full head of another was replaced, to ensure the photos were consistently categorized into one of the three treatment groups. Cheek employed a screening process to select an audio recording of a choir that was considered to exhibit a "racially neutral" vocal quality, meaning that listeners were unable to discern the performers' racial identity based on sound alone. This process was somewhat compromised by the need to also find a recording that adequately represented the same-sized choir as in the photographs (Cheek, 2007, pp. 45–48), and a decision was taken to select a recording with a better fit in terms of chorister numbers but scored lower on the "racial neutrality" measure. Twenty-six American choral music educators (15 Black, 11 White) were divided into the three treatment groups ($n = 10$, $n = 8$, and $n = 8$) and asked to rate the performance of the choir using an adaption of the *California Music Educators Association Bay Section Vocal Adjudication Form*. Adjudicators were subsequently

asked to complete McCrary's (1993) *Racial-Encounter Measure* online in their own time, so that evaluations might be compared with evaluators' general racial attitudes.

An ANOVA showed a main effect of treatment, with a significant difference between the highest average ratings given to the homogeneous Black choir, and the lowest average ratings to the homogeneous White choir (Cheek, 2007, p. 72). There were no significant differences between the ratings of Black and White adjudicators across the conditions and no interactions between adjudicator race and the treatment conditions (Cheek, 2007, p. 80). No correlations with racial attitudes could be made, due to low compliance with this aspect of the study. Cheek speculated that the low compliance may have been due both to participants perceiving the underlying purpose of the study once they received the post-treatment survey and the possible perception that the survey (originally designed to be administered to teenagers) was not relevant to their experiences.

Cheek's doctoral study raises important questions around the consensual racialization of research subjects, the racialized characteristics of musical performances, and potential racial bias in the evaluative frameworks in music education contexts to be explored further. However, it is the only study to date to attempt to capture characteristics of evaluators beyond race and gender categories.

VanWeelden's (2004) and VanWeelden and McGee's (2007) studies examined the effect of racially stereotyped music and the perception of conductor race on ratings of conductor and choral ensemble performance in the United States. In the 2004 study, 6 male graduate student and faculty member choral conductors were recruited (3 Black, 3 White) and, controlling for location and dress, were video-recorded conducting to a 45-s audio recording of an arrangement of the African American spiritual, "Ezekiel Saw de Wheel". This piece was specifically selected because it is historically from Black culture but is now performed by choirs of all ages and ethnicities. Conductors completed the task several times and recordings were selected of each conductor to match the others in gesture, eye contact, facial expression, and posture. The six selected conductor recordings were partially randomized into six stimulus videotapes with short excerpts of Renaissance music interleaved between each conductor's video to help create a "renewed listening slate". One hundred and sixty-nine American undergraduate music students across six universities (69 male, 100 female; 66 vocalists, 103 instrumentalists—racial identities are not reported) were asked to rate the performance on qualities commonly used for choral festival adjudications: ensemble intonation, tone quality, attacks and releases, phrasing, dynamics, balance and blend, and diction. Evaluators were also asked to rate each conductor's eye contact, facial expression, posture, and overall effectiveness, as well as their confidence in the conductor.

An ANOVA found main effects for conductor race and tape order, with ensemble performance ratings for the three Black conductors (taken together) significantly higher than for White conductors (VanWeelden, 2004, pp. 41–42). Moderately to moderately strong correlations were found between ensemble performance ratings and evaluators' ratings for conductor body expressions (eye contact, facial expression, and posture), effectiveness, and evaluators' confidence (VanWeelden, 2004, p. 44).The latter finding was corroborated by Morrison et al.'s (2009) research, which found that the level of conductors' gestural expressivity had a significant effect on the evaluation of ensemble expressivity. VanWeelden acknowledged that it is not possible for this study alone to conclude whether the combination of racially stereotyped music and the perception of conductor race affected the performance evaluations of the ensemble and of the conductors themselves, noting the possibility that in this case the Black conductors were simply better than their White colleagues.

To investigate this further, VanWeelden and McGee (2007) explored the effect of repertoire racially stereotyped as either Black or White on the evaluation of an ensemble's performance believed to be conducted by conductors perceived as Black or White. Modeled on the 2004 study, 4 professional male conductors (2 Black, 2 White) were video-recorded, each conducting to a 45-s recorded excerpt of the spiritual "Ezekiel Saw de Wheel" (1 Black, 1 White) or a piece of Western art music: Felix Mendelssohn's "When God Commanded Angels" (1 Black, 1 White), performed by the same choir (racial identities not specified). They aimed to control for location, attire, racially stereotyped diction (e.g., changing the word "de" to "the" in the spiritual), and again selecting video recordings of conductors that were best matched for gesture and facial expression. Three hundred and fifty-three American undergraduate music students (252 Black, 101 White) viewed one of four partially randomized stimulus videotapes, each showing a Black conductor conducting the spiritual, a White conductor conducting the spiritual, a Black conductor conducting the piece of Western art music, and a White conductor conducting the piece of Western art music. The same evaluation form was as used in the 2004 study, and Renaissance music was again interleaved.

An ANOVA revealed a significant interaction between repertoire and conductor's race. Ensemble performance ratings were significantly higher when Black conductors were perceived to be conducting the spiritual and White conductors were perceived to be conducting the piece of Western art music with no effect from evaluators' own racial identities (i.e., no "in-group bias") (VanWeelden & McGee, 2007, p. 11). Significant interactions were also found for eye contact, facial expression, and posture, which were rated more highly for Black conductors when perceived to be conducting the spiritual and White conductors when perceived to be conducting the piece of Western art music, again with no effect from evaluators' own racial identities (VanWeelden & McGee, 2007, pp. 13–14). VanWeelden and McGee observed that the lack of in-group bias present in their study differed from the

findings of earlier research by Killian (1990) and Morrison (1998), which showed that Black and White students preferred performers that were the same race as themselves (VanWeelden & McGee, 2007, p. 15). It also differs from Davidson and Edgar's analysis as described earlier.

In summary, these studies collectively reveal the complexity of racialized bias in HECM performance evaluations and the different ways it manifests in specific musical contexts. All seven studies presented in this article reported a significant effect of race in performance evaluations as well as significant interactions with evaluator race and gender (Clauhs, 2013; Davidson & Edgar, 2003), performer instrument (Elliott, 1995), repertoire performed (Clauhs, 2013; VanWeelden & McGee, 2007), performer movement (Peynircioğlu et al., 2018), and performer gender (Clauhs, 2013; Davidson & Edgar, 2003; Elliott, 1995; Peynircioğlu et al., 2018). This work underscores the real-world impact of performance evaluations—appraisals with consequences for educational outcomes and career progressions. All of the studies recognized the importance of the work, but noted that their findings could not be generalized and called for further research.

As further investigations are designed, it is worth reflecting on several issues that these studies and their write-ups raise. In the following section I discuss four areas in turn: the practice of racialization and the choice of race categories and terminology; defining and controlling for racialized characteristics of/associations with performance variables and the use of racially biased evaluative frameworks; the perpetuation of harmful stereotypes in study designs and results interpretations; and the use of evaluator characteristics in study designs.

Issues Arising from These Studies

Racialization and the Choice of Race Categories

Establishing the perceived racialized group of research subjects and participants as well as describing them with race terminology was fundamental to these studies' designs. Racialization of research subjects without their consent for use in study designs and non-consensual racialization of participants for the purposes of analysis (as distinct from racial self-identification) was variable throughout the studies. In some instances, researchers clearly stated whether participants (evaluators) self-identified (Cheek, Peynircioğlu et al., VanWeelden & McGee), and in other places, non-consensual racialization was suggested through context. Only in Cheek's study was it made explicit that the models used for the stimulus self-identified their racial categories.

Many of these studies relied on assumptions about how people are racialized to control for the variable of race. Cheek addresses this in his doctoral dissertation, and Clauhs does too to a limited degree, but of the peer-reviewed studies, only VanWeelden and McGee included a discussion about race as a social concept and the role of visual information in racialization in order to justify their study design. They stated that:

> A person's skin color and other physical traits (e.g., nose structure) may not be fully indicative of his or her race […] Still, for the purpose of this current study […] the term 'race' will be used since it is often the conclusion drawn, however correctly or incorrectly, from a first impression of certain physical features, and frequently affects subsequent perceptions. (VanWeelden & McGee, 2007, p. 8)

Peynircioğlu et al.'s occlusion of head and hands and Cheek's changing of skin tone, hair color, and a model's head to control for the variable of race in their study, speaks to the prominent role of these assumptions in their research designs.

Martinez (2023) offers a detailed critique of this practice in perceptual studies more broadly, labeling it "facecraft", a form of Fields and Fields' (2012) concept of "racecraft". Racecraft practices, he explains, perpetuate racist ideologies by reinforcing the notion of race; in the case of facecraft, through the false belief that biological traits define racial categories, which are then imposed on individuals. Martinez argues that such "research practices often fall short of illuminating the operations of racism (and often obfuscate them) because they themselves constitute a form of racecraft" (Martinez, 2023, p. 1). In his article, he underscores the multifaceted and inherently subjective process of racialization, highlighting that researchers cannot assume how participants will racialize a subject (undermining the reliability of the variable for the study design), and drawing attention to the potential bias present even in existing image banks designed for this purpose. Rather than preclude the use of physical traits in perceptual studies, he calls for greater care in study designs and clarity in language, noting "the important difference between discussing 'the white faces' versus 'the faces with [list of shared facial features] perceived to be white by X% of the sample'" (Martinez, 2023, p. 187).

In interaction with this is the choice of race categories that were used by researchers as the assumptive racialized groups. None of the researchers provided a reason for their use of specific race categories and their relevance for their research context. A partial exception to this was Peynircioğlu et al., who explained that "although only East Asians appear to be subjected to this bias, we will keep the overall 'Asian' referent to be consistent with the literature" (Peynircioğlu et al., 2018, p. 296); however, the specific literature to which they were referring was not made clear. In the detail of the study's design it is revealed that they are using the term "Asian" to refer specifically to performers that are "ostensibly from China, Korea, Japan and Vietnam" (Peynircioğlu et al., 2018, p. 302). Davidson and Edgar use the more specific racial categories of Afro-Caribbean, White European, and Indian Asian, but offer no explanation as to why they were chosen for the study's design or if the research subjects self-identified as

such. Throughout the write-up they elide these more geographically specific racial categories with Black and White, but without explanation for this choice. There is a sense of "we all know who we mean" inherent in these study designs. This lack of clarity regarding variable definition, beyond itself reinforcing unhelpful racialized generalizations, undercuts the value of the findings.

Greater care might also be taken to understand the historical basis for specific race categories, before employing them as variable descriptors. For example, the racial category Caucasian was used by Clauhs, Peynircioğlu et al., and VanWeelden and McGee. While still commonly used to describe those of White European heritage (particularly within the United States census), it actually originates from a eugenicist taxonomy for the "ideal type of human" (Mukhopadhyay, 2008, p. 13). As mentioned in my introduction, how race categories are understood to be more or less acceptable changes over time, but due diligence is of course paramount in this context.

Finally, the additional slippage between the language of "race", "culture", and "ethnicity" in Peynircioğlu et al.'s study is an indication of the lack of reflection and indeed specificity about their use of race as a variable in their work. More problematically, this slippage suggests that socially complex aspects of a person's background such as "culture" and "ethnicity" (Suyemoto et al., 2020) can be inferred from names or physical features, one of the ways that—perhaps inadvertently—such write-ups can reproduce harmful stereotypes. Suyemoto et al. (2020) note that in empirical research contexts, mixing lay meanings of terminology with formal discipline-specific definitions "threatens valid operationalizations for research and challenges our ability to use these concepts in interventions to promote social justice and psychological health" (Suyemoto et al., 2020, p. 1). Clauhs acknowledges that a significant design limitation is that only four individuals were standing in for entire populations and individual skills and characteristics undoubtedly impacted the evaluations to some degree, a point VanWeelden also noted. This is essentially true for all the studies, and a challenge for future research designs to navigate.

Racialized Characteristics of/Associations with Study Variables and the Potential Bias of Evaluative Frameworks

Attention to the racialized associations of musical materials and therefore the notion of "fit" with the perceived racialized identity of performers was handled differently in the study designs. Researchers engaged with the issue directly as the focus of their studies (Clauhs and VanWeelden & McGee,), attempted to control for it (Cheek), or left it unaddressed (Elliott, Davidson & Edgar, Peynircioğlu et al., and VanWeelden). At the core of this notion of fit is what appears to be a persistent belief that musicians only have an innate capacity for music of their supposed cultural heritage (Wang, 2014, p. 94; see also Leppänen, 2015; Thurman, 2019). An anecdote from American violinist Jennifer Koh (born of Korean parents) provides a clear example:

> In the beginning of my career, I was told by an influential conductor—who had never heard me play—that I could never be a true artist because he did not understand Chinese music and therefore Chinese people could never understand classical music. (Koh, 2021)

The interaction of racialized musical fit with race and gender was particularly striking in Clauhs's study, with Black males associated most strongly with jazz music and yet receiving lower scores than White males.

Similarly, racialized performance characteristics may also need to be taken into consideration. Cheek attempts to control for racial neutrality in the aural stimulus but reveals that the performers of the overdubbed stimulus are a racially White choir. This raises the question as to how the aural stimulus employed in the study was perceived to represent a racially neutral performance in terms of style or vocal characteristics when it was actually performed by a White choir. Cheek recounts that a Black female participant, an experienced high school choral music director, expressed her surprise and admiration upon hearing a Black high school choral group deliver a performance matching the level of competency demonstrated in the recording (Cheek, 2007, p. 89). Cheek speculated that it was possible this view was held by other subjects as well —that "the adjudicators […] may have anticipated hearing vocal qualities in the performance of the choral group they perceived to be Black that are inconsistent with prevailing standards for choral performance" (Cheek, 2007, p. 89). Put another way, and considering the higher scores given to the Black choir by both Black and White adjudicators, Cheek's interpretation of the comment suggests that when a choir visually perceived to be homogeneously Black sounds like a homogenously *White* choir, they are perceived to be exceeding expectations. The racial identities of the performers of the recordings used in the overdubbed conditions of all other studies were not reported. This observation is not meant to suggest that there is or is not an inherent relationship between vocal sound and a performer's racial identity (see Eidsheim (2015) for a detailed discussion on this), but that such factors need careful consideration in stimulus creation.

Related to this is the ongoing concern with the racial biases of evaluative instruments. These studies underscore that further reflection is needed on how evaluative frameworks and concepts (e.g., "excellence"), tools (e.g., music competition scoring protocols), and privileging of performance modes (e.g., "sound" over other performance elements) may be racially biased, or even a product of "white racial framing" (Ewell, 2020, after Feagin, 2013) within music education and the HECM sector more widely. While it may not always be feasible to develop

new evaluative tools and validate them within the resources or timeframe of a research project, continued interrogation and contextualization of evaluative frameworks is clearly needed.

Perpetuation of Stereotypes

It is standard practice for researchers to provide a discussion of their findings, which often includes some speculation on the factors that may have contributed to the results. However, there is a risk of perpetuating harmful stereotypes in this process, as these studies reveal. For example, a theory of low expectation of b/Black performers (similar to Cheek's above) was also put forth by two other studies to explain their results. Elliott suggested that "low teacher expectation" of minority students may have been the reason that the Black student performers were scored significantly lower than the White students (Elliott, 1995, p. 53). Davidson and Edgar present the same theory but in positive terms for White pianists as mentioned earlier in this article, explaining that evaluators' preference for White pianists' performances might be explained by evaluators' preconceptions that White pianists are generally better than their b/Black counterparts. The fact that the same explanation has been used to justify both higher and lower scores for b/Black performers compared to their White counterparts underscores the limited utility of such conjecture.

Davidson and Edgar's study also appeared to normalize racial in-group bias and engage in unnecessary speculations which have the potential to reinforce harmful stereotypes. As one example, the authors reported significant difference in the rating patterns between Judge demographic categories (male/female, Afro-Caribbean/White European) and Modes of Performance, noting that male judges gave a wide spread of marks across the different performance modes (full view, point-light display, sound only, etc.) while females' scores differentiated between performance modes very little. White European females gave high scores in all but the sound-only condition while Afro-Caribbean females gave consistently lower scores. Noting that there were many possible interpretations of these results, they offered two: "(i) The women may be more cautious than the men, tending to rate the performers more consistently, with Black women being the least confident in their judgments; (ii) The low sound-only ratings could reflect the necessity for visual cues" (Davidson & Edgar, 2003, p. 176).

If the findings had been reversed, with White European men's ratings more consistent across the Modes of Performance, would this have been attributed to their lack of confidence? It could have been equally speculated whether the Afro-Caribbean women's scores represented more consistent critical evaluation of the performances regardless of Performance Mode, and an indication that they were less influenced by other nominally extrinsic factors such as gesture when evaluating the performances. Such an assertion would be equally as unhelpful in unpicking stereotypes and biases. The authors also speculate on the "necessity" of visual cues for the White European females in their evaluations, which both suggests a privileging by the researchers of the aural over the embodied performance event, and something of a contradiction in terms of the critique of Afro-Caribbean women's ratings.

Evaluator Characteristics

In advocating for the use of controlled experiments to research racial discrimination, Blank et al. propose that one of their major contributions is that they have the capability to identify the "characteristics of people who are more or less likely to exhibit discriminatory attitudes and behaviours" (Blank et al., 2004, p. 6). Cheek's attempt to find a correlation between evaluators' scores and more general racial attitudes by using McCrary's *Racial-Encounter Measure* was the only study to attempt to obtain characteristics of evaluators beyond race and gender, such as attitudes, beliefs, socio-economic and ethnic backgrounds, professional role, education and training, or even mood. As further research is undertaken, the characteristics of research participants including the societal roles they represent (teachers, adjudicators, agents, administrators, and other gatekeepers) will be an important focus. Intercultural contexts that recognize the multiplicity of engagement with HECM worldwide will be critical to examine, as well as the musical contexts of small and large ensemble settings with their profiles of performer-to-performer power relations (conductors, section leaders, rank-and-file players) and the prestigious positions that ensembles like string quartets and orchestras hold within the global HECM industries.

Conclusions: The Call for More Interdisciplinary Collaboration

This collection of findings reveals that, much like discrimination in everyday life, racial discrimination in HECM music performance evaluations emerges through interaction with multiple intersecting factors and manifests differently depending on the specific musical and cultural contexts. Controlled experimental studies have been identified as an essential tool in investigating racial discrimination (Blank et al., 2004) and add an important dimension to existing theoretical and qualitative research (e.g., Bradley, 2016; Bull 2019; Thurman, 2019, Yang, 2007). For example, findings from controlled experimental designs can be a powerful tool in helping to confront ongoing debates about whether or not the low representation of racially minoritized performers in the HECM industry is a "pipeline" issue or the result of nuanced racial bias and discrimination in the evaluation of musical performances.

To design, conduct, and report research in this area that does not intentionally or unintentionally rely on or affirm stereotypes, expose research subjects to harmful racialization or

stereotyping, or impact the career prospects of real-world musicians, requires detailed and thoughtful study designs devised in collaboration with interdisciplinary partners. These include researchers in music perception and cognition, performance studies, music sociology, and the sociology of discrimination, specialists in the philosophy of music aesthetics, and performers with lived experiences of discrimination in HECM musical contexts. Heightened attention to the nuances of power dynamics in decision-making processes, and how research collaborators are remunerated and appropriately recognized for their contributions, will be an important component, as well as a critical awareness of the colonial histories in which this research takes place (Jacoby et al., 2020).

There is great scope for innovation in this area of music research with increased institutional support for interdisciplinary collaboration, new technological tools quickly emerging for the control of the crucial variables of music performance (deepfakes, avatars, extended reality, etc.), and the ability to work with colleagues across the globe easier than ever before. The knowledge and resources needed to overcome the social and technical challenges necessary to undertake this research are already available in our collective musical community, the questions just need to be prioritized.

Action Editor

Emily Payne and Karen Burland, University of Leeds, School of Music

Peer Review

David John Baker, University of Amsterdam, Institute for Logic, Language, and Computation
Francesca Carpos, Guildhall School of Music and Drama, Institute for Social Impact Research in the Performing Arts

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval

This research did not require ethics committee or IRB approval. This research did not involve the use of personal data, fieldwork, or experiments involving human or animal participants, or work with children, vulnerable individuals, or clinical populations.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Cayenna Ponchione-Bailey https://orcid.org/0000-0001-9348-3221

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Notes

1. The first author was approached for clarification during the preparation of this article but by the time of publication had not yet been able to respond with an explanation.
2. See Schimmack (2021) and Kurdi et al. (2021), for recent debates about the IAT's construct validity and its usefulness as a tool for measuring racial bias.

References

Baker, D. J., Belfi, A., Creel, S., Grahn, J., Hannon, E., Psyche, L., Margulis, E. H., Schachner, A., Schutz, M., Shanahan, D., & Vuvan, D. T. [Society for Music Perception and Cognition Anti-Racism and Equity Committee and Executive Board] (2020). Embracing anti-racist practices in the music perception and cognition community. *Music Perception*, *38*(2), 103–105. https://doi.org/10.1525/mp.2020.38.2.103

Behne, K. E., & Wöllner, C. (2011). Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication. *Musicae Scientiae*, *15*(3), 324–342. https://doi.org/10.1177/1029864911410955

Bermingham, G. A. (2000). Effects of performers' external characteristics on performance evaluations. *Update: Applications of Research in Music Education*, *18*(2), 3–7. https://doi.org/10.1177/875512330001800202

Bland, M., & Cho, E. (2021). The effect of physical movement on observers' perception of musical quality in a choral performance. *Psychology of Music*, *49*(6), 1449–1461. https://doi.org/10.1177/0305735620959424

Blank, R. M., Dabady, M., & Citro, C. F. (Eds.) (2004). *Measuring racial discrimination*. The National Academies Press. https://doi.org/10.17226/10887

Bradley, D. G. (2016). Hidden in plain sight: Race and racism in music education. In C. Benedict, P. Schmidt, G. Spruce, & P. Woodford (Eds.), *The Oxford handbook of social justice in music education* (online edition, pp. 190–203). Oxford University Press.

Brown, B. K. (2020a). When Black conductors aren't comfortable at concerts, classical music has a real problem. *Level*. Retrieved January 31, 2023, from https://level.medium.com/black-concert-trauma-5fa0459e5b3

Brown, D. (2020b). An open letter on racism in music studies: Especially ethnomusicology and music education. *My People Tell Stories*. Retrieved January 28, 2023, from https://www.mypeopletellstories.com/blog/open-letter

Bull, A. (2019). *Class, control, and classical music*. Oxford University Press. https://doi.org/10.1093/oso/9780190844356.001.0001

Cheek, J. A. (2007). *The effect of race and racial perception on adjudicators' ratings of choral performances attributed to racially homogeneous and racially heterogeneous groups* [Doctoral dissertation, The University of North Carolina at Greensboro]. NC Docks. https://libres.uncg.edu/ir/uncg/f/umi-uncg-1308.pdf

Chmurzyńska, M. (2023). Research on the assessment of music performance from Maria Manturzewska's early experiments to more recent studies. *Musicae Scientiae*, *27*(4), 862–874. https://doi.org/10.1177/10298649231188272

Clauhs, M. S. (2013). *The effects of race and gender bias on style identification and music evaluation* [Doctoral dissertation, Temple University]. TUScholarShare. https://scholarshare.temple.edu/bitstream/handle/20.500.12613/981/Clauhs_temple_0225E_11379.pdf

Cox, T., & Kilshaw, H. (2021). *Creating a more inclusive classical music: A study of the English orchestral workforce and the current routes to joining it*, Executive summary. Arts Council England. Retrieved May 2, 2022, from https://www.artscouncil.org.uk/sites/default/files/download-file/Executive_Summary.pdf

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, *21*(2), 103–113. https://doi.org/10.1177/030573569302100201

Davidson, J. W., & Edgar, R. (2003). Gender and race bias in the judgement of western art music performance. *Music Education Research*, *5*(2), 169–181. https://doi.org/10.1080/1461380032000085540

Di Laccio, G. (2022). *Equality and diversity in global repertoire: Orchestras season 2021–2022*. Donne, Women in Music. Retrieved August 15, 2023, from https://donne-uk.org/wp-content/uploads/2021/03/Donne-Report-2022.pdf

Doeser, J. (2016). *Racial / ethnic and gender diversity in the orchestra field.* League of American Orchestras. Retrieved January 29, 2023, from http://www.ppv.issuelab.org/resources/25840/25840.pdf

Douglas, A. (2021, May 7). *A critique of ethnicity taxonomies.* AD Blog. Retrieved August 5, 2023, from https://alexanderdouglas.info/2021/05/07/a-critique-of-ethnicity-taxonomies

Eagle, O. M., Mendoza, J. K., Wolf, J. E., Baker, D. J., & Vuvan, D. T. (2023). Anti-racism and equity panel: How can music science be more socially just? *Auditory Perception & Cognition*, *6*(3–4), 369–386. https://doi.org/10.1080/25742442.2023.2236540

Eidsheim, N. S. (2015). Race and the aesthetics of vocal timbre. In O. Bloechl, M. Lowe, & J. Kallberg (Eds.). *Rethinking difference in music scholarship* (pp 338–365). Cambridge University Press. https://doi.org/10.1017/CBO9781139208451.012

Elliott, C. A. (1995). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education* (The 15th International Society for Music Education: ISME Research Seminar), *127*, 50–56. https://www.jstor.org/stable/40318766

Ewell, P. A. (2020). Music theory and the white racial frame. *Music Theory Online*, *26*(2), 59–87. https://doi.org/10.30535/mto.26.2.4

Feagin, J. R. (2013). *The white racial frame: Centuries of racial framing and counter-framing* (2nd ed.). Routledge.

Fields, K. E., & Fields, B. J. (2012). *Racecraft: The soul of inequality in American life*. Verso.

Flagg, A. (2020). Anti-Black discrimination in American orchestras. *Symphony Magazine.* Summer 2020, 30–37. Retrieved May 2, 2022, from https://americanorchestras.org/wp-content/uploads/2021/01/Anti-Black-Discrimination-in-American-Orchestras.pdf

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae*, *17*(3), 265–276. https://doi.org/10.1177/1029864913495404

Gittens, I., Ddungu, R., Stevens, H., Beaumont, C., & Wilson, R. (2021). *Being Black in the UK music industry*. Black Lives in Music. Retrieved January 29, 2023, from https://blim.org.uk/report/report-9cpolsaz9uja/

Griffiths, N. K. (2008). The effects of concert dress and physical appearance on perceptions of female solo performers. *Musicae Scientae*, *12*(2), 273–290. https://doi.org/10.1177/102986490801200205

Griffiths, A. (2020). Playing the white man's tune: Inclusion in elite classical music education. *British Journal of Music Education*, *37*(1), 55–70. https://doi.org/10.1017/S0265051719000391

Harrington, A. M. (2018). The effect of implied performer age and group membership on evaluations of music performances. *Update: National Association of Music Education*, *36*(2), 5–12. https://doi.org/10.1177/8755123317725726

Hochman, A. (2019). Racialization: A defense of the concept. *Ethnic and Racial Studies*, *42*(8), 1245–1262. https://doi.org/10.1080/01419870.2018.1527937

Jacoby, N., Margulis, E. H., Clayton, M., Hannon, E., Honing, H., Iversen, J., Klein, T. R., Mehr, S. A., Pearson, L., Peretz, I., Perlman, M., Polak, R., Ravignani, A., Savage, P. E., Steingo, G., Stevens, C. J., Trainor, L., Trehub, S., Veal, M., & Wald-Fuhrmann, M. (2020). Cross-cultural work in music cognition: Challenges, insights, and recommendations. *Music Perception*, *37*(3), 185–195. https://doi.org/10.1525/mp.2020.37.3.185

Kawabata, M. (2023). The new "yellow peril" in "western" European symphony orchestras. In A. Bull, C. Scharff, & L. Nooshin (Eds.), *Voices for change in the classical music profession: New ideas for tackling inequalities and exclusions* (pp. 159–171). Oxford University Press. Retrieved August 15, 2023, from https://doi.org/10.1093/oso/9780197601211.003.0015

Kendi, I. X. (2019). *How to be an anti-racist*. The Bodley Head; Random House; One World.

Killian, J. N. (1990). Effect of model characteristics on musical preference of junior high students. *Journal of Research in Music Education*, *38*(2), 115–123. https://doi.org/10.2307/3344931

Koh, J. (2021). A violinist on how to empower Asian musicians. *The New York Times*, 21 July 2021. Retrieved January 29, 2023, from https://www.nytimes.com/2021/07/21/arts/music/jennifer-koh-asians-classical-music.html

Kolbe, K. (2022). Producing (musical) difference: Power, practices and inequalities in diversity initiatives in Germany's classical music sector. *Cultural Sociology*, *16*(2), 231–249. Retrieved August 15, 2023, from https://doi.org/10.1177/17499755211039437

Kroger, C., & Margulis, E. H. (2017). "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music*, *45*(1), 49–64. https://doi.org/10.1177/0305735616642543

Kurdi, B., Ratliff, K. A., & Cunningham, W. A. (2021). Can the implicit association test serve as a valid measure of automatic cognition? A response to Schimmack (2021). *Perspectives on*

Psychological Science, *16*(2), 422–434. https://doi.org/10.1177/1745691620904080

Leppänen, T. (2015). The west and the rest of classical music: Asian musicians in the Finnish media coverage of the 1995 Jean Sibelius Violin Competition. *European Journal of Cultural Studies*, *18*(1), 19–34. https://doi.org/10.1177/1367549414557804

Martinez, J. E. (2023). Facecraft: Race reification in psychological research with faces. *Journal of Vision (Charlottesville, Va.)*, *23*(9), 182–194. https://doi.org/10.1167/jov.23.9.4673

McCrary, J. (1993). Effects of listeners' and performers' race on music preferences. *Journal of Research in Music Education*, *41*(3), 200–211. https://doi.org/10.2307/3345325

Morrison, S. J. (1998). A comparison of preference responses of white and African-American students to musical versus musical/visual stimuli. *Journal of Research in Music Education*, *46*(2), 208–222. https://doi.org/10.2307/3345624

Morrison, M. D. (2019). Race, blacksound, and the (re)making of musicological discourse. *Journal of the American Musicological Society*, *72*(3), 781–823. https://doi.org/10.1525/jams.2019.72.3.781

Morrison, S. J., Price, H. E., Geiger, C. G., & Cornacchio, R. A. (2009). The effect of conductor expressivity on ensemble performance evaluation. *Journal of Research in Music Education*, *57*(1), 37–49. http://www.jstor.org/stable/40204947

Morrison, S. J., Price, H. E., Smedley, E. M., & Meals, C. D. (2014). Conductor gestures influence evaluations of ensemble performance. *Frontiers in Psychology*, *5*, Article 806. https://doi.org/10.3389/fpsyg.2014.00806

Mukhopadhyay, C. C. (2008). Getting rid of the word 'Caucasian'. In M. Pollack (Ed.), *Everyday antiracism: Getting real about race in school* (pp. 12–16). The New Press.

Nápoles, J., Silvey, B. A., & Montemayor, M. (2021). The influences of facial expression and conducting gesture on college musicians' perceptions of choral conductor and ensemble expressivity. *International Journal of Music Education*, *39*(2), 260–271. https://doi.org/10.1177/0255761420926665

Nguyễn, A. N., & Pendleton, M. (2020, March 23). *Recognizing race in language: Why we capitalize "Black" and "White"*. Center for the Study of Social Policy. Retrieved January 29, 2023 from https://cssp.org/2020/03/recognizing-race-in-language-why-we-capitalize-black-and-white/

Nooshin, L. (2011, December). Introduction to the special issue: The ethnomusicology of western art music. *Ethnomusicology Forum*, *20*(3), 285–300. https://doi.org/10.1080/17411912.2011.659439

Nusseck, M., & Wanderley, M. M. (2009). Music and motion–how music-related ancillary body movements contribute to the experience of music. *Music Perception*, *26*(4), 335–353. https://doi.org/10.1525/mp.2009.26.4.335

Peynircioğlu, Z. F., Bi, W., & Brent, W. (2018). The "Asian bias" illusion in musical performance: Influence of visual information. *The American Journal of Psychology*, *131*(3), 295–305. https://doi.org/10.5406/amerjpsyc.131.3.0295

Pistorius, J. (2019). Predicaments of coloniality, or, opera studies goes ethno. *Music & Letters*, *100*(3), 529–539. https://doi.org/10.1093/ml/gcz046

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709

Sauvé, S. A., Phillips, E., Schiefelbein, W., Daikoku, H., Hegde, S., & Moore, S. (2023). Anti-colonial strategies in cross-cultural music science research. *Music Perception*, *40*(4), 277–292. https://doi.org/10.1525/mp.2023.40.4.277

Scharff, C. (2015). Equality and diversity in the classical music profession. King's College London. Retrieved July 9, 2021, from https://www.impulse-music.co.uk/wp-content/uploads/2017/05/Equality-and-Diversity-in-Classical-Music-Report.pdf

Schimmack, U. (2021). Invalid claims about the validity of implicit association tests by prisoners of the implicit social-cognition paradigm. *Perspectives on Psychological Science*, *16*(2), 435–442. https://doi.org/10.1177/1745691621991860

Spence, S. (2021, April). *Equity, diversity and inclusion: A research report exploring workforce diversity and representation in London music education hubs through the lens of racism.* Music Mark. Retrieved May 2, 2022, from https://www.musicmark.org.uk/wp-content/uploads/Music-Mark-EDI_Report.pdf

Sussman, R. W. (2014). *The myth of race*. Harvard University Press. https://www.jstor.org/stable/j.ctt9qdt73

Suyemoto, K. L., Curley, M., & Mukkamala, S. (2020). What do we mean by "ethnicity" and "race"? A consensual qualitative research investigation of colloquial understandings. [special issue, P. J. Aspinall & C. Caballero (Eds.). *Genealogies of racial and ethnic representation*]. *Genealogy (Basel)*, *4*(3), 81. https://doi.org/10.3390/genealogy4030081

Tan, S. E. (Ed.) (2021). Decolonising music and music studies [Special edition]. *Ethnomusicology Forum*, *30*(1). https://doi.org/10.1080/17411912.2021.1938445

Thurman, K. (2019). Performing lieder, hearing race: Debating blackness, whiteness, and German identity in interwar central Europe. *Journal of the American Musicological Society*, *72*(3), 825–865. https://doi.org/10.1525/jams.2019.72.3.825

Vann, M. (2018). Lack of diversity in top orchestras remains a major challenge for musicians of color. *NBC News Online*. Retrieved January 31, 2023, from https://www.nbcnews.com/news/us-news/lack-diversity-top-orchestras-remains-major-challenge-musicians-color-n891386

VanWeelden, K. (2004). Racially stereotyped music and conductor race: Perceptions of performance. *Bulletin of the Council for Research in Music Education*, *160*, 38–48. https://www.jstor.org/stable/40319217

VanWeelden, K., & McGee, I. R. (2007). The influence of music style and conductor race on perceptions of ensemble and conductor performance. *International Journal of Music Education*, *25*(1), 7–17. https://doi.org/10.1177/0255761407074886

Vuoskoski, J. K., Thompson, M. R., Spence, C., & Clarke, E. F. (2016). Interaction of sight and sound in the perception and experience of musical performance. *Music Perception*, *33*(4), 457–471. https://doi.org/10.1525/mp.2016.33.4.457

Wang, G. (2014). *Soundtracks of Asian America: Navigating race through musical performance*. Duke University Press. https://doi.org/10.1215/9780822376088

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, *45*(3), 470–479. https://doi.org/10.2307/3345540

Wapnick, J., Mazza, K., & Darrow, A.-A. (1998). Effects of performer attractiveness, stage behaviour and dress on violin performance evaluation. *Journal of Research in Music Education*, *46*(4), 510–521. https://doi.org/10.2307/3345347

Williams, K., & Bain, V. (2022). *Dignity at work 2: Discrimination in the music sector*. Incorporated Society of Musicians. Retrieved August 15, 2023, from https://www.ism.org/images/files/ISM-Dignity-2-report.pdf

Yang, M. (2007). East meets west in the concert Hall: Asians and classical music in the century of imperialism, post-colonialism, and multiculturalism. *Asian Music*, *38*(1), 1–30. https://www.jstor.org/stable/4497039

Yoshihara, M. (2021). *Anti-Asian discrimination in American orchestras.* League of American Orchestras. Retrieved August 15, 2023, from https://americanorchestras.org/wp-content/uploads/2021/11/04_Fall-2021-Anti-Asian-Discrimination-in-American-Orchestras.pdf