

## rijksuniversiteit groningen

# Anti-Female Gender Stereotypes and Gender Bias in the Assessment of Music 

Carla Stella Finsterbusch<br>Master Thesis - Talent Development and Creativity

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.


#### Abstract

As females are underrepresented in the music industry and male artists dominate the music charts, this research aims to advance our understanding of these gender differences. Prior research hinted at gender stereotypes and gender bias against music by female artists. If antifemale bias is discovered, it needs to be reduced, at least in formal music assessments such as exams and awards. This study investigated (1) whether there are negative stereotypic beliefs against music by female artists, (2) whether music presented as made by female artists is judged less favorably than the same music presented as made by male artists, and (3) whether this potential anti-female bias can be reduced in formal music assessment by using more standardized assessment procedures. A $2 \times 2$ mixed experiment $(N=89)$ was conducted, in which artist gender (male versus female, within-factor) and assessment approach (holistic versus mechanical, between-factor) were manipulated, and electronic music tracks were assessed. Contrary to expectations, this study did not find negative stereotypes against music by female artists, and music by female artists was not rated significantly lower than music by male artists so the effect of structured, mechanical assessment to reduce anti-female gender bias could not be supported per definition. Due to the insignificant and small effects, which were partly opposite the hypothesized direction, pointing toward the possibility of positive stereotypes about music by female artists, no firm conclusions could be drawn, and future research incentives are discussed.


Keywords: music assessment, anti-female stereotypes, anti-female bias, holistic versus mechanical assessment

## Anti-Female Gender Stereotypes and Gender Bias in the Assessment of Music

Despite a trend toward increased streaming of female artists, females are continuously underrepresented in the music industry (Hayes, 2019; Smith et al., 2021). Between 2012 and 2020, females represented less than one-third of artists, approximately $12.6 \%$ of songwriters, and less than 3\% of music producers of popular songs (Smith et al., 2021). There also seems to be a listening preference for male artists as female artists only made up approximately onefourth of top Spotify, Apple Music, and Billboard charts from 2017 to 2019 (Hayes, 2019), and songs by male artists tend to stay in the charts longer (Carter, 2018). This gender imbalance calls for an exploration of underlying factors.

Although traditional gender stereotypes are quite persistent over time, maintaining discriminatory behavior (Castillo-Mayén \& Montes-Berges, 2014; Haines et al., 2016), very little research on gender stereotypes and gender bias has been conducted in the domain of music. Furthermore, the few existing studies focused on one single genre, had low power, showed mixed results, might be outdated, or did not keep music performance constant across artist gender so that assessment differences may have resulted from actual performance differences and not from gender bias (see e.g., Colley et al., 2003; Davidson \& Edgar, 2003; Millar, 2008; North et al., 2003). Therefore, the first aim of this study is to investigate whether negative stereotypes are held against music by female artists and whether music presented as made by a female artist is judged less favorably than the same music presented as made by a male artist.

Music is not only judged by music listeners but also in formal assessment contexts such as music exams, competitions, and award shows, which require a more objective procedure to ensure unbiased evaluation (Saunders \& Holahan, 1997). Differential evaluation of female and male music artists in formal assessment contexts may indirectly influence streaming and listening behavior in an informal context. For example, artists who were

Grammy winners sold more records due to the promotion of their popularity (Watson \& Anand, 2006). While the default, intuitive way of assessing music tends to be biased, assessing music performance in a structured and mechanical way might reduce such bias by focusing attention on relevant information (see e.g., Dawes et al., 1989; Lenney et al., 1983; Milkman et al., 2009; Wolgast et al., 2017). Thus, the second aim of this study is to investigate if potential anti-female gender bias in formal music assessment can be prevented or reduced by assessing music in a more standardized way.

## Literature Review

## Gender Stereotypes Against Female Music Artists

As stereotypic gender beliefs can result in biased judgment, it is crucial to investigate the potential influence of gender stereotypes attributed to music artists (Heilman \& ParksStamm, 2007). Using gender stereotypes, people make inferences about others' personal attributes grounded upon gender-based social categories (Ashmore \& Del-Boca, 1979; Eagly \& Wood, 2012). As stereotypes tend to be triggered automatically, they pose a danger for evaluation (Heilman \& Parks-Stamm, 2007). Consequently, the lower ranking of female artists in music charts suggests that they may be perceived differently (Carter, 2018).

In line with the Social Role Theory, Eagly and Wood (2012) argued that people's inferences about gender attributes are rooted in evolutionary sex differences and manifest in gender role beliefs performed as social roles. They explained that according to the correspondent inference principle, people tend to infer dispositions of men and women when observing their behavior, thereby shaping gender role beliefs (Eagly \& Wood, 2012).

Relating the correspondent inference principle to the music industry, it may be expected that observing the dominance of male artists in the music industry may induce people to assume that being a musician is a more male-typed gender role and that men typically have the relevant attributes to be musicians, perhaps especially in genres that are more dominated by
male artists such as R\&B, hip-hop, rock, and electronic music (Carter, 2018). Accordingly, traditional gender stereotypic attributes were inferred differently from music composers thought to be male or female, even though the music presented was the same (North et al., 2003). Moreover, male participants were found to hold negative gender stereotypes against female composers (North et al., 2003). Further, stereotypic beliefs regarding the instruments boys and girls should practice were revealed (Davidson \& Edgar, 2003). In addition, while a preference for classical music was negatively related to using gender stereotypes, a preference for electronic dance music, hip-hop, or hard-house music was positively related to using gender stereotypes (Ter Bogt et al., 2010). As this study focused on electronic music, the existence of stereotypes was expected to be more likely. This study aimed to bridge the gap in research by investigating whether people indeed hold negative stereotypic beliefs against music by female artists. Therefore I hypothesize:

Hypothesis 1: Negative stereotypic beliefs are held against music by female artists.
Prior research did not provide clear expectations or theories about differences between male and female judges' stereotypic beliefs about male and female artists. While male and female artists were found to be judged differently by male and female assessors, and anti-female stereotypes were shown in male participants (see e.g., Colley et al., 2003; Millar, 2008; North et al., 2003), this study explored whether male and female participants hold different stereotypic beliefs.

## Anti-Female Gender Bias in Music Assessment

The persistent gender imbalance in the music charts suggests that music by women tends to be undervalued by music listeners (Colley et al., 2003). As stereotypes were shown to influence preferences (Ter Bogt et al., 2010), gender stereotypes may negatively bias music judgment. According to the Lack of Fit Model, information that is inconsistent with stereotype-based expectations tends to be interpreted according to those expectations so that
judgment becomes negatively biased for gender role inconsistent behavior (Heilman \& ParksStamm, 2007). As music listeners may perceive the music industry as male-gender typed, they may develop negative performance expectations towards female artists and may consequently interpret their music less favorably (Carter, 2018; Heilman \& Parks-Stamm, 2007; Smith et al., 2021). Since behavioral ratings tend to be more influenced by expectations than by actual memories of an event, music by female artists may be judged less favorably according to their negative stereotype-based expectations (Heilman \& Parks-Stamm, 2007).

While anti-female gender bias has been detected in other judgment contexts, its complexity needs to be investigated in more depth (Swim et al., 1989). In the context of music, male artists' creations were found to be preferred by music listeners and recognized as higher in worth and artistic merit than female artists (Millar, 2008). Likewise, male New Age music composers were rated higher on musical competence than female composers (Colley et al., 2003). Further, an anti-female bias was discovered in male students' judgments of jazz excerpts of various music composers (North et al., 2003). Besides this, performance judgments were found to be influenced by perceptions of the instrument as male or female gender-typed (Elliot, 1996). Moreover, the extent of gender bias was found to be related to the perception of a genre as male- or female-typed: Female artists were rated less favorably in a male-typed genre (North et al., 2003). This study focused on electronic music, a generally male-typed genre (Smith et al., 2021). On the other hand, studies revealed a pro-female bias in music judgment, particularly in female judges (Davidson \& Edgar, 2003; North et al., 2003). Moreover, a more favorable evaluation of female composers was observed when more information, such as a short biography, was provided as compared to only the composer's name (Colley et al., 2003).

Besides the above explained mixed findings discovering both anti-female and antimale bias (see e.g., Colley et al., 2003; Davidson \& Edgar, 2003; North et al., 2003), most
research had substantial methodological flaws. For example, Colley and colleagues (2003) only investigated between-subject differences for artist gender, but not within-subject differences, and the study had low power and may not generalize to genres other than New Age music. Similarly, Davidson and Edgar's (2003) study had low power and only allowed for conclusions related to Western Art music. As Millar's (2008) study only collected qualitative data about musical preferences, artist gender was not manipulated, and hence, the study did not allow for conclusions about actual gender bias. The present study aimed to improve on these flaws by conducting an experiment keeping music constant across artist gender and checking for between- and within-subject differences in music judgment in order to investigate whether music by female artists is judged less favorably than music by male artists. I hypothesize:

Hypothesis 2: Music presented as made by female artists is judged less favorably than the same music presented as made by male artists.

Prior research indicated a different extent and direction of bias for male and female assessors. For example, male had a stronger pro-male bias in music preferences than females (Millar, 2008). Similarly, male assessors were biased against female artists in judgments of jazz excerpts, whereas a pro-female bias was found in judgments by female assessors (North et al., 2003). Moreover, females tend to give higher ratings overall (Colley et al., 2003). This study additionally explored the effect of participant gender on the possible gender bias.

## Gender Bias Reduction Through Structured and Mechanical Assessment

If the intuitive judgments of music indeed result in a gender bias against music by female artists, it is critical to explore how this bias can be reduced or prevented. Unless people are motivated to process information systematically to increase accuracy, they tend to process information by relying on stereotypes to save cognitive resources (Chaiken, 1980; Heilman \& Park-Stamm, 2007). As stereotypes guide people's selective attention toward cues
indicating social category and result in gender role expectations aligning with these social categories, tools to direct attention away from cues are needed (Ashmore \& Del-Boca, 1979; Eagly \& Wood, 2012). As attitudes tend to change if behavior changes, a decrease in biased judgment may contribute to an attitude change and diminish gender stereotypes (Lenney et al., 1983). In order to identify ways that reduce the potential anti-female gender bias in music judgment, it is essential to take a closer look at how music is assessed in formal assessment settings.

While holistic assessment reflects the assessor's intuition, in mechanical assessment, criteria and predictors are determined, which are used to collect relevant information, which is then combined and weighted using a decision rule (Dawes et al., 1989; Meijer et al., 2020; Neumann et al., 2021; Wolgast et al., 2017). Whereas holistic judgment is often biased and results in self-fulfilling prophecies, mechanical assessment diminishes bias, is fairer, and improves performance predictions, thereby increasing the validity and reliability of assessment (Dawes et al., 1989; McPherson \& Thompson, 1998; Milkman et al., 2009; Neumann et al., 2021; Wolgast et al., 2017). By directing attention toward specific performance-related information (Lenney et al., 1983) and challenging expectations related to gender stereotypes (Eagly \& Wood, 2012), structured, mechanical assessment may also reduce the potential anti-female bias in music assessment.

Accordingly, criteria-specific rating scales were found to contribute to the reliability of woodwind and brass performance judgments (Saunders \& Holahan, 1997). Whereas male assessors neither had an anti-female gender bias when using clear evaluation guidelines nor when using ambiguous guidelines, female assessors did have an anti-female bias when using ambiguous assessment guidelines, but not when using clear guidelines (Lenney et al., 1983). Nevertheless, this finding indicates that a mechanical approach in music assessment may
reduce anti-female gender bias by increasing attention toward relevant musical aspects. Therefore, I hypothesize:

Hypothesis 3: Using a structured, mechanical assessment approach decreases antifemale gender bias in music assessments.

As more structure in assessment can be increased in various ways (Wolgast et al., 2017), the different effects of using no structure, rating structure, or a decision rule on a decrease in anti-female gender bias in music assessments were additionally explored. When using a rating structure, information is combined intuitively after using specific criteria as a guideline, whereas a decision rule mechanically combines the information gathered through criteria by assigning weights to criteria (Meijer et al., 2020; Wolgast et al., 2017). Besides this, as research suggested that male and female assessors may judge music by female and male artists differently in holistic and mechanical assessment (Lenney et al., 1983), the effect of participant gender on the reduction of the potential bias was explored.

Overall, investigating the research questions at hand ultimately aimed to contribute to a more comprehensive understanding of potential influencing factors for the underrepresentation of females in the music industry.

## Method

A $2 \times 2$ mixed experiment was conducted, in which assessment procedure (betweensubjects factor: holistic vs. mechanical) and artist gender (within-subjects factor: male vs. female) were manipulated. Participants rated instrumental electronic music tracks, which were chosen as tracks with vocals would not allow for manipulating artist gender.

## Participants

Based on an a priori power analysis, a sample size of $n=128$ was required to attain a medium effect size (Cohen's $f=.25$ ) and a power of .80 . In total, 128 participants took part in this study. Three participants that did not complete the entire survey, two participants not
identifying as male or female, 12 extreme outliers on mean music ratings for male artists and mean music ratings for female artists in holistic and mechanical assessment as well as on mean stereotype ratings (i.e., more than two standard deviations from the mean) and 22 respondents who seemed to have responded carelessly were excluded. Hence, analyses were conducted with a sample of $n=89\left(76.4 \%\right.$ females, $23.6 \%$ males, $\left.M_{\text {age }}=20, S D_{\text {age }}=1.81\right)$. Careless responses were identified through counterbalanced items. As the first three items in the stereotypes questionnaire included one counterbalanced item, scoring $1=$ strongly disagree on at least all first three items indicated inconsistency and a high likelihood of careless responding.

Extreme outliers may have represented legitimate individual differences in stereotype and music ratings, and participants may have been incorrectly identified as careless (Meijer et al., 2016). As the sample was much smaller $(n=89)$ and had lower power, separate analyses were conducted with data from 123 participants, including extreme outliers and careless responses to check for deviations (see Appendix A for demographics). Participants were firstyear psychology bachelor students, Dutch and international, at the University of Groningen, voluntarily participating for course credit. Participation required an understanding of English.

## Research Design and Procedure

A 2 (artist gender: male versus female; within-factor) x 2 (assessment approach: holistic versus structured/mechanical; between-factor) mixed experiment was conducted, with music ratings as the dependent variable. The study was approved by the ethics committee of the Faculty of Behavioral and Social Sciences at the University of Groningen under the name "PSY-2122-S-0193 Musical Preferences and the Judgement of Music". The experiment was administered in English, using the online software Qualtrics. Students could complete the survey between April 22 and May 30, 2022. Participation was voluntary, and participants were compensated with course credit. While the estimated response time was 30
minutes, and participants were allowed to pause and resume the survey, the midpoint of response time in this sample $(n=89)$ was only eleven minutes $\left(M d n_{\text {response-time }}=11 \mathrm{~min}\right.$, $\left.S D_{\text {response-time }}=1749 \mathrm{~min}\right)$.

After providing basic demographic information, each participant listened to nine 30second audio fragments of electronic music tracks without vocals. As performance judgments are usually made within 15 to 30 seconds, 30 seconds were deemed an adequate time to make the first judgment of music (Colley et al., 2003; Davidson \& Edgar, 2003). The average listening time for each track was recorded in seconds $\left(M_{\text {listeningtime }}=32 \mathrm{~s}, S D_{\text {listeningtime }}=11 \mathrm{~s}\right.$ ). One male and one female fictitious artist profile were created for each of the nine tracks so that the within-factor artist gender could be manipulated (see Appendix B). In order to increase the credibility of the artist profiles, each profile included a picture indicating the artist's gender and brief background information about their age, city/country of origin (all popular European countries and their capitals), and their first name, which was based on a Google search about popular gender-typed first names in the respective city/country of origin. Corresponding male and female artist profiles attributed to the same track differed in their artist picture, name, and gender, but matched in city/country of origin and age.

Participants were randomly allocated to either the holistic ( $n=46$ ) or the structured/mechanical group $(n=43)$ and listened either to four tracks presented as made by female artists and five tracks presented as made by male artists, or vice versa (see Appendix C). Tracks were presented in a randomized order, and after listening to and rating a track, they could move on to the next one. The holistic group provided an overall rating of their track liking, while the structured/mechanical group rated each track on seven different criteria as well as one item assessing overall track liking after rating the criteria. One track was repeated, but the second time, it was presented with the artist profile of the opposite gender to check whether the same participant rated music presented as made by a female artist
differently than the same music presented as made by a male artist. Lastly, participants filled out a questionnaire assessing stereotypic beliefs regarding music by female and male artists. At the end of the survey, they were debriefed about the study's actual purpose.

## Measures

## Holistic Assessment of Music

The holistic group rated their overall liking of the audio clip on a 7-point Likert scale ("How much did you like the track you just listened to?"; $1=$ dislike a great deal; $7=$ like a great deal) based on previous research on gender bias in the evaluation of music (Colley et al., 2003; North et al., 2003) (see Table D1).

## Structured/Mechanical Assessment of Music

The structured/mechanical group rated each clip on 7-point Likert scales for the seven criteria rhythm, mood/atmosphere, innovativeness, interestingness ${ }^{1}$, expressiveness, technical competence, and artistic merit (e.g., "I think the track is rhythmical."; $1=$ strongly disagree; 7 = strongly agree) (see Table D1). Those criteria were based on previous research about gender bias in music evaluation (Colley et al., 2003; North et al., 2003) and on nine components regarded as essential for good music (Millar, 2008). After rating the track on the seven criteria, they indicated their overall liking of the track as done by the holistic group.

## Gender Stereotypes of Male and Female Music Artists

Based on the seven assessment criteria, a questionnaire investigated stereotypic gender beliefs about music by female and male artists (e.g., "Music by male artists is usually more interesting than music by female artists"; $1=$ strongly disagree; $7=$ strongly agree) (see Table D2). The three items assessing gender stereotypes regarding innovativeness, artistic merit, and expressiveness of music by female and male artists were counterbalanced.

[^0]
## Data Analysis Strategy

The analyses were run using the software SPSS Statistics, version 26.0.0.1. Hypotheses were tested using a mixed analysis of variance, investigating between (assessment approach) and within (artist gender) factors. First, a one-sample t-test investigated Hypothesis 1, testing whether participants' scores were significantly higher than the scale midpoint of four, given that higher scores indicated more anti-female stereotypes. Counterbalanced items were recoded accordingly. A two-sample t-test was used to explore the relationship of participant gender with gender stereotypes.

In order to analyze differences in music ratings for female and male artists, mean rating scores were computed separately for the liking of female artists and liking of male artists. Mean calculations excluded the track that was repeated; separate mean ratings for the liking of female and male artists were computed for the repeated track. Mean ratings when using no structure were based on the overall track liking of the holistic group, mean ratings when using a decision rule were based on the structured/mechanical group's average of ratings for all seven, equally-weighted assessment criteria, and mean ratings when using rating structure were based on the structured/mechanical group's overall track liking.

Hypothesis 2 was investigated using a 2 (artist gender) x2 (assessment approach: holistic vs. mechanical) mixed ANOVA with music ratings as the dependent variable, and the main effect of artist gender was evaluated. Further, a $2 \times 2 \times 2$ mixed ANOVA adding participant gender as a second between-factor to the model explored the main effect of participant gender and the interaction effect between participant gender and artist gender.

A second way to investigate Hypothesis 2 was to analyze the main effect of artist gender for ratings by the same participant of the repeated track that was presented once as made by a female artist and another time as made by a male artist using a separate 2 (artist gender) x2 (assessment approach: holistic vs. mechanical) mixed ANOVA with music ratings
as the dependent variable. Further, a $2 \times 2 \times 2$ mixed ANOVA including participant gender as between-factor explored the interaction effect between artist gender and participant gender.

Hypothesis 3 investigated the interaction effect between artist gender and assessment approach in the same 2 (artist gender) x2 (assessment approach: holistic vs. mechanical) mixed ANOVA used to analyze Hypothesis 2. The $2 \times 2 \times 2$ mixed ANOVA with participant gender as a second between-factor explored the interaction effect between artist gender, participant gender, and assessment approach. The different effects of using no structure, rating structure, or a decision rule on a decrease in the potential anti-female gender bias were explored by inspecting the interaction effect between artist gender and assessment approach in a separate 2 (artist gender: male vs. female) x2 (assessment approach: holistic vs. structured) mixed ANOVA with music ratings as the dependent variable.

Differences between holistic and mechanical ratings of the repeated track that was presented to each participant were investigated with the interaction effect between artist gender and assessment approach in the respective $2 \times 2$ (assessment approach: holistic vs. mechanical) mixed ANOVA. Also, the interaction effect between artist gender, assessment approach, and participant gender was explored in the respective $2 \times 2 \times 2$ mixed ANOVA. Further, differences between holistic and structured ratings of the repeated track were explored by inspecting the interaction effect between artist gender and assessment approach in the respective $2 \times 2$ (assessment approach: holistic vs. structured) mixed ANOVA.

## Results

## Assumptions

The assumptions that need to be met for conducting a mixed ANOVA were checked. QQ-plots showed slight deviations from normality, but this assumption was considered as met, as ANOVA is quite robust against normality violations in subgroup samples larger than 25 (Van den Berg, n.d.-a), which was the case for the within-factor artist gender and the
between-factor assessment approach. Only the subgroup of male participants used for the exploratory analyses was smaller than $25(n=21)$. Violations of homogeneity, which were present in this study as indicated by Levene's test results, were considered unproblematic as within- and between-factor subgroup sample sizes were roughly equal (Van den Berg, n.d.-a). Only the subgroup samples for the between-factor participant gender were unequal due to the predominantly female sample. Hence, results for the $2 \times 2 \times 2$ ANOVA adding participant gender as between-factor may have been misleading. For the one-sample $t$-test, the independence assumption was met, and a QQ-plot indicated that the dependent variable was approximately normally distributed. As explained above, 12 extreme outliers and 22 participants identified as careless respondents were excluded ( $n=89$ ), but analyses were additionally conducted with those participants included $(n=123)$ in order to compare differences in results (see Appendix G for results when including outliers and careless responses).

## Descriptive Statistics

Descriptive statistics are shown in Figure 1. In addition, separate music ratings for female and male artists given by male and female participants for the three assessment approaches are displayed in Appendix E for both the main sample excluding outliers and careless responses and the sample including those cases. Further, Figure F illustrates the separate mean ratings for male and female participants when excluding outliers and careless responses.

## Figure 1.

Mean Music Ratings of Music Perceived as Made by Female and Male Artists for Holistic, Structured, and Mechanical Assessment


Note. $N=89$. Holistic condition: $n=46$, mechanical condition: $n=43, \mathrm{~F}=$ female artist, M $=$ male artist, ratings were on 7-point Likert scales ranging from $1=$ dislike a great deal to 7 $=$ like a great deal, error bars: +/- 2 SD.

## Hypothesis Testing

## Anti-Female Gender Stereotypes

Investigating if people held negative stereotypic beliefs against music by female artists (Hypothesis 1) showed a mean $(M=3.7, S D=.41)$ in the opposite direction than hypothesized. Hence, significance testing was redundant, and Hypothesis 1 could not be supported. For completeness, a one-sample t-test explored stereotypic beliefs in the opposite direction; mean stereotypes ratings were significantly lower than the scale midpoint with a
moderate to large effect size $(t(88)=85.21, p<.001, d=.73)$. A similar finding was shown when including extreme outliers and careless responses (see Table G1).

A two-sample t-test exploring the effect of participant gender on the extent of gender stereotypes against female artists showed no significant anti-female stereotypic beliefs; mean stereotype ratings were not significantly higher for male participants ( $M=3.82, S D=.30$ ) than for female participants $(M=3.66, S D=.43)$ and the effect size was small $(t(88)=1.61$, $p=.11, d=.43$ ). In contrast, this effect was significant with a medium effect size when including outliers and careless responses (see Table G2).

## Anti-Female Gender Bias

Testing the effect of artist gender on music ratings (Hypothesis 2), a $2 \times 2$ mixed ANOVA did not reveal a significant effect and the effect size was small $(F(1)=00, p=.98$, $\left.\eta_{p}{ }^{2}=.00\right)$; music ratings for female artists $(M=4.37, S D=.72)$ were not significantly lower than those for male artists ( $M=4.38, S D=.77$ ). Thus, Hypothesis 2 was not supported. In contrast, an analysis without outliers and careless responses found higher music ratings for female artists (see Table E2), but the main effect was small and not significant (see Table G3). Despite the insignificant result for anti-female bias, a $2 \times 2 \times 2$ mixed ANOVA conducted as planned neither found a significant main effect for participant gender $(F(1)=2.85, p=.1$, $\eta_{p}{ }^{2}=.03$ ), nor a significant interaction effect between participant gender and artist gender $\left(F(1)=1.51, p=.22, \eta_{p}{ }^{2}=.02\right.$ ); while female participants gave lower music ratings to female artists $(M=4.4, S D=.74)$ than to male artists $(M=4.45, S D=.73)$, but male participants gave higher music ratings to female artists $(M=4.28, S D=.62)$ than to male artists ( $M=$ 4.15, $S D=.86$ ), female and male participants' music ratings for male and female artists did not differ significantly differently. In contrast, when including outliers and careless responses, male and female participants both gave significantly higher music ratings to female artists than to male artists, and the difference between music ratings for male and
female artists was significantly larger in male participants, but effect sizes were small (see Table E2 for mean ratings and Table G4 for results).

Opposite to expectations, investigating differences in ratings of the repeated track assessed by the same participant discovered higher music ratings for female artists ( $M=4.64$, $S D=1.01)$ than for male artists $(M=4.47, S D=1.02)$. Exploring gender bias in the opposite direction with a $2 \times 2$ mixed ANOVA found no significant main effect for artist gender, and the effect size was small $\left(F(1)=3.24, p=.08, \eta_{p}{ }^{2}=.04\right)$. In contrast, including outliers and careless responses yielded significantly higher music ratings for female artists but a small effect size (see Table E4 for mean ratings and Table G5 for results). Testing the effect of participant gender on differences between music ratings of the repeated track with a $2 \times 2 \times 2$ mixed ANOVA showed a small and not significant interaction effect between artist and participant gender $\left(F(1)=.02, p=.89, \eta_{p}{ }^{2}=.00\right)$; female and male participants did not rate music by female and male artists significantly differently (see Table E3 for mean ratings). When including outliers and careless responses, findings coincided with each other (see Table E4 for mean ratings and Table G8 for results).

## Reduction of Anti-Female Gender Bias Through Mechanical Assessment

Given that no anti-female bias was found, Hypothesis 3 can no longer be adequately investigated. However, for completeness of the planned analyses, a $2 \times 2$ mixed ANOVA was conducted and found no significant difference between music ratings for female and male artists when using no structure (holistic assessment) as compared to a decision rule (mechanical assessment) $\left(F(1)=2.36, p=.13, \eta_{p}{ }^{2}=.03\right)$; female artists were given lower music ratings than male artists in holistic assessment (female artists: $M=4.2, S D=.79$; male artists: $M=4.29, S D=.89$ ), and higher ratings than male artists in mechanical assessment (female artists: $M=4.56, S D=.58$; male artists: $M=4.48, S D=.71$ ). Including outliers and careless responses did not show a significant effect either (see Table G3), but female artists
received higher music ratings in both holistic and mechanical assessment (see Table E2). Further, a $2 \times 2 \times 2$ mixed ANOVA exploring the difference between male and female participants' music ratings for male and female artists in holistic and mechanical assessment revealed a small and not significant interaction effect $\left(F(1)=2.06, p=.16, \eta_{p}{ }^{2}=.02\right)$; while female participants' holistic ratings were lower for female artists $(M=4.24, S D=.81)$ than those for male artists ( $M=4.4, S D=.76$ ) and their mechanical ratings were higher for female artists $(M=4.6, S D=.6)$ than those for male artists ( $M=4.51, S D=.7$ ), male participants' holistic ratings were higher for female artists $(M=3.97, S D=.65)$ than those for male artists $(M=3.75, S D=.92)$ and, likewise, their mechanical ratings were higher for female artists ( $M$ $=4.47, S D=.54)$ than for male artists $(M=4.4, S D=.75)$. When including outliers and careless responses, the interaction effect was significant with a small to moderate effect size (see Table E2 for mean ratings and Table G4 for results). For completeness, a 2x2 mixed ANOVA additionally explored whether music ratings for female and male artists differed when assessing without structure (holistic assessment) as compared to with rating structure (structured assessment) and showed a small and not significant interaction effect $(F(1)=$ $2.39, p=.13, \eta_{p}{ }^{2}=.03$ ); while music ratings were lower for female artists when using no structure, they were higher for female artists than for male artists when using rating structure (see Table E1). While the interaction effect was not significant either when including outliers and careless responses (see Table G6), female artists were rated higher in both holistic and structured assessments (see Table E2).

Although music ratings for the repeated tracks presented to the same participant were higher for female artists, the differences in those ratings when using no structure (holistic) as compared to a decision rule (mechanical) were explored for completeness by a $2 \times 2$ mixed ANOVA and were found to be small and not significant $\left(F(1)=.06, p=.81, \eta_{p}{ }^{2}=.00\right)$; both holistic and mechanical music ratings for female artists were higher than holistic and
mechanical music ratings for male artists (see Table E3). When including outliers and careless responses, a similar result was found (see Table E4 for mean ratings and Table G5 for results). Further, a $2 \times 2 \times 2$ mixed ANOVA did not find significant differences between female and male participants' music ratings of the repeated track for male and female artists in holistic and mechanical assessment, and the effect size was small $\left(F(1)=.05, p=.82, \eta_{p}{ }^{2}\right.$ $=.00$ ) (see Table E3 for mean ratings). A similar result was found when including outliers and careless responses (see Table E4 for mean ratings and Table G8 for results). Besides this, a $2 \times 2$ mixed ANOVA exploring differences between music ratings of the repeated track for female and male artists when using no structure (holistic) as compared to rating structure (structured) found a small and not significant interaction effect $\left(F(1)=.05, p=.82, \eta_{p}{ }^{2}=\right.$ .00); structured music ratings for female artists were higher than both, structured and holistic music ratings for male artists, but lower than holistic ratings for female artists (see Table E3). Including outliers and careless responses revealed a similar result (see Table E4 for mean ratings and Table G7 for results).

## Discussion

This study aimed to contribute to a more comprehensive understanding of potential influencing factors for the persistent gender imbalance in the music industry by investigating the extent of negative gender stereotypes and gender bias against music by female artists and of structured/mechanical assessment to reduce this potential bias. It was hypothesized that people hold negative stereotypic beliefs against music by female artists, that music presented as made by female artists is judged less favorably than the same music presented as made by male artists and that structured/mechanical assessment reduces the potential anti-female gender bias.

Unexpectedly, the results of this study did not reveal significantly negative stereotypes against music by female artists (Hypothesis 1). Surprisingly, results for this
sample may even indicate a trend toward negative stereotypes against music by male artists. This finding contrasts prior research revealing anti-female gender stereotypes in male assessors (North et al., 2003). While results may even point toward positive stereotypes for music by female artists, at this point, there is no theoretical rationale explaining this claim, and further research is needed to make firm conclusions.

Surprisingly, in this study, ratings for music presented as made by female artists were not significantly lower than those for male artists (Hypothesis 2). Interestingly, opposite to expectations, testing differences in ratings for the repeated track given by the same participant indicated that music by female artists might be rated slightly higher, but not significantly more favorably, than music by male artists, and thus pointed towards a similar effect as shown when testing Hypothesis 1, which may hint at a tendency towards a possibly more positive evaluation of female artists and a more negative evaluation of male artists. Indeed, prior research revealed an anti-male bias in music assessment (Davidson \& Edgar, 2003; North et al., 2003). The findings in this study contradict prior research showing that music by female artists is judged less favorably (Colley et al., 2003; Millar, 2008; North et al., 2003). Further, contradicting prior research suggesting that males tend to have a stronger anti-female bias (Millar, 2008; North et al., 2003), participant gender did not influence music ratings for male and female artists significantly differently in this study. However, as effects were insignificant and small, and prior research presented similarly mixed findings, no robust conclusions can be drawn, and further research is needed.

As no anti-female gender bias was found, the hypothesis that structured, mechanical assessment reduces bias against music by female artists (Hypothesis 3 ) could not be supported per definition. Exploring the potentially differential effect between using no structure, rating structure, and decision rules may possibly indicate a less favorable judgment of female artists than of male artists when using no structure (holistic assessment), but
perhaps only concerning ratings by female assessors, and a more favorable judgment of female artists when using rating structure (structured assessment) and when using decision rules (mechanical assessment). However, no firm conclusions can be drawn as no anti-female gender bias was detected in the first place, findings were not significant, and effect sizes were small. These findings challenge prior literature indicating that more structure in assessment procedures improves selection decisions and increases validity and reliability and that male artists are more likely to have an anti-female bias (McPherson \& Thompson, 1998; Millar, 2008; Neumann et al., 2021; North et al., 2003; Wolgast et al., 2017). However, on the grounds of the insignificant findings for a bias against music by female artists, the null hypothesis should not be accepted, and further research should be conducted.

Results may be due to socially desirable responses (Baumeister \& Vohs, 2007). Besides this, the information provided about artists in the artist profiles may have diminished the extent of gender bias (Colley et al., 2003; North et al., 2003). Further, as participants were students in the Netherlands, a generally progressive country valuing gender equality, social role changes may have diminished their use of gender stereotypes (Castillo-Mayén \& Montes-Berges, 2014; European Institute for Gender Equality, n.d.). Lastly, the deviations between results when excluding outliers and careless responses compared to results when including those cases may indicate that they have indeed distorted the results, and highlight the need for further investigation.

## Limitations and Future Directions

## Limitations Due to Sample Characteristics

Some limitations of this study were due to the characteristics of the sample at hand. First, the sample consisted mainly of female participants, which may have distorted the results as females were previously shown to give higher ratings overall, and anti-female bias tends to be more prevalent in male assessors (Millar, 2008; North et al., 2003). However,
while exploring the effect of participant gender did not show significant differences between their music ratings for female and male artists, it pointed towards the possibility that female, not male participants, may have an anti-female bias. Second, as participants were psychology students and not well-versed in the field of music assessment, and as no real formal music assessment procedure was investigated, results are not generalizable to a more diverse population or formal music assessment contexts. Third, while the short response time for the study may indicate that the participants may have merely been motivated to participate to earn credits, resulting in careless responding, the time needed to complete the survey may have been overestimated. At least, checking the average listening time per track indicated that, on average, they listened to tracks at their entire length.

## Limitations Due to Study Design

Other limitations may have been due to the study design. First, this study strictly only allows for conclusions about electronic music without vocals, and it is unknown if the results generalize to other genres. However, keeping music constant across artist gender would have been challenging when using music with vocals. Second, although counterbalanced items detected careless responses, there may have been alternative explanations for those responses. Third, the study's purpose may have been too obvious, and due to the self-report format, socially desirable responses may have been an issue (Baumeister \& Vohs, 2007). Participants may have biased future participants by passing on information about the true research questions after participating and reading the debriefing form. Fourth, while the information provided in the fictitious artist profiles may have influenced participants' music ratings (Colley et al., 2003), this was expected to only have a negligible influence as music was kept constant across artist gender and artist profiles for the same song were at least kept similar in terms of age and origin. Next, while the seven assessment criteria items were taken from other studies investigating gender bias in music, their quality is unclear, and they may have
been ambiguous (Heilman \& Parks-Stamm, 2007). Further, an order effect may have influenced ratings of the mechanical group as assessment criteria were always shown in the same order (McPherson \& Thompson, 1998).

## Limitations Due to Analyses

Further limitations concerned the conducted analyses. As most effect sizes were small, those effects may have only explained a small proportion of the variance in the sample (Van den Berg, n.d.-b). Besides, while this study focused on the results when excluding outliers and careless responses and therefore had a loss of power, it is uncertain if including or excluding those cases provides the most representative data. There may have been insufficient power to find significant differences in ratings for female and male artists. Further, while a limitation of the analyses of differences for music ratings of the same track presented to the same participant twice was that it was based on only one track, having multiple repeated tracks may have been too obvious. Moreover, as the normality and homogeneity assumptions were violated in the $2 \times 2 \times 2$ mixed ANOVA investigating the effect of participant gender, these results may be misleading, and no firm conclusions could be drawn from the respective results. Nevertheless, the effect size (i.e., partial eta squared) may be construed despite those violations (Van den Berg, n.d.-a).

As the hypotheses were not supported, and in light of this study's shortcomings as well as the contradiction of its results in context with other studies, no substantial practical or theoretical implications can be drawn without further investigation. Nevertheless, this study provides valuable suggestions for future research.

## Methodological Considerations for Future Research

Future research should replicate this study with a sample with a broader age range, a balance between male and female participants, and participants from different backgrounds; ideally, experts from the music industry to check whether results may differ. Moreover,
susceptibility to impression management should be reduced through, for example, direct observations of music listeners' judgment. Further, conducting a study in a formal music assessment setting is crucial as judgments in this context may indirectly affect the streaming of award winners (Watson \& Anand, 2006). Further, as music ratings vary depending on the amount of information provided about the artist (Colley et al., 2003; North et al., 2003), potential differences in results if only the artist's gender is revealed should be explored. Additionally, potential differences in the extent of anti- female bias in other music genres should be examined.

## Critical Topics for Future Research on Gender Stereotypes and Bias in Music Judgment

As the primary aim of this research was to investigate factors that may potentially influence the underrepresentation of females in the music industry, future research should continue to scrutinize these gender differences and examine other potentially influential factors involved. Besides this, it should explore if female but not male assessors show an anti-female bias, if there is an anti-female bias in holistic but not in mechanical assessment, and if there may even be negative stereotypes and a bias against music by male artists. Further, the degree of evidence for the null hypothesis of no differences between ratings for music female and male artists using Bayesian analyses should be investigated

## Conclusions

Taken together, this study neither provided support for negative gender stereotypes against music by female artists nor for a gender bias against music by female artists. While social desirability could have influenced results, participants may have held little negative performance expectations against female artists, and other factors may influence the underrepresentation of female artists. However, results may be different in a formal music assessment context. Importantly, due to the unexpected and null findings, no firm implications can be formulated. Notwithstanding its limitations, this study contributes
valuable new research impulses and draws attention to the importance of further clarification of potential reasons behind female artists' underrepresentation in the music industry.

## References

Ashmore, R. D., \& Del Boca, F. K. (1979). Sex stereotypes and implicit personality theory: toward a cognitive-social psychological conceptualization. Sex Roles: A Journal of Research, 5(2), 219-248. https://doi.org/10.1007/BF00287932

Carter, B. (2018). Male musical artists still dominate the charts - and the margin isn't small. CentralSauce. Retrieved November 11, 2021, from https://centralsauce.com/gender-diversity-music-charts

Castillo-Mayén, R., \& Montes-Berges, B. (2014). Analysis of current gender stereotypes. Anales De Psicología, 30(3), 1044-1060. https://doi.org/10.6018/analesps.30.2.138981

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology, 39(5), 752-766. https://doi.org/10.1037//0022-3514.39.5.752

Colley, A., North, A., \& Hargreaves, D. J. (2003). Gender bias in the evaluation of new age music. Scandinavian Journal of Psychology, 44(2), 125-131. https://doi.org/10.1111/1467-9450.00330

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

Dawes, R. M., Faust, D., \& Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668-1674. https://doi-org.proxy-ub.rug.nl/10.1126/science. 2648573

Eagly, A. H., \& Wood, W. (2012). Social role theory. In P. van Lange, A. Kruglanski, \& E. T. Higgins (Eds.), Handbook of theories in social psychology (pp. 458 - 476). Thousand Oaks, CA: Sage Publications. http://dx.doi.org/10.4135/9781446249222.n49

Elliott, C. A. (1996). Race and gender as factors in judgments of musical performance. Bulletin of the Council for Research in Music Education, 127(127), 5056.

European Institute for Gender Equality. (n.d.). Gender equality index. European Institute for Gender Equality. Retrieved June 25, 2022, from https://eige.europa.eu/gender-equality-index/2021/country/NL

Haines, E. L., Lofaro, N., \& Deaux, K. (2016). The times they are a-changing ... or are they not? A comparison of gender stereotypes, 1983-2014. Psychology of Women Quarterly, 40(3), 353-363. https://doi.org/10.1177/0361684316634081

Hayes, J. (2019). The gender play gap: Breaking down men's \& women's performance on music charts. Chartmetric. Retrieved November 11, 2021, from https://blog.chartmetric.com/the-gender-play-gap/

Heilman, M. E., \& Parks-Stamm, E. J. (2007). Gender stereotypes in the workplace: Obstacles to women's career progress. In S.J. Correll (Eds.). Social psychology of gender (Ser. Advances in group processes, v. 24) (pp. 44-77). Jai Press Inc.

Lenney, E., Mitchell, L., \& Browning, C. (1983). The effect of clear evaluation criteria on sex bias in judgments of performance. Psychology of Women Quarterly, 7(4), 313328. https://doi.org/10.1111/j.1471-6402.1983.tb00846.x

McPherson, G., \& Thompson, W. (1998). Assessing music performance: Issues and influences. Research Studies in Music Education, 10(1), 12-24. https://doi.org/10.1177/1321103X9801000102

Meijer, R. R., Neumann, M., Hemker, B. T., \& Niessen, S. (2020). A tutorial on mechanical decision-making for personnel and educational selection. Frontiers in Psychology, 10, [3002]. https://doi.org/10.3389/fpsyg.2019.03002

Meijer, R. R., Niessen, A. S. M., \& Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. Assessment, 23(1), 52-62. https://doi.org/10.1177/1073191115577800

Milkman, K. L., Chugh, D., \& Bazerman, M. H. (2009). How can decision making be improved? Perspectives on Psychological Science, 4(4), 379-383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Millar, B. (2008). Selective hearing: Gender bias in the music preferences of young adults. Psychology of Music, 36(4), 429-445. https://doi.org/10.1177/030573560708604

Neumann, M., Niessen, A. S. M., \& Meijer, R. R. (2021). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. Organizational Psychology Review, 11(3), 205239. https://doi.org/10.1177/2041386620983419

North, A. C., Colley, A. M., \& Hargreaves, D. J. (2003). Adolescents' perceptions of the music of male and female composers. Psychology of Music, 31(2), 139-154. https://doi.org/10.1177/0305735603031002291

Saunders, T. C., \& Holahan, J. M. (1997). Criteria-specific rating scales in the evaluation of high school instrumental performance. Journal of Research in Music Education, 45(2), 259-272. https://doi.org/10.2307/3345585

Smith, S. L., Pieper, K., Choueiti, M., Hernandez, K., \& March, K. Y. (2021). Inclusion in the recording studio? Gender and racelethnicity of artists, songwriters \& producers across 900 popular songs from 2012-2020. USC Annenberg Inclusion Initiative. https://assets.uscannenberg.org/docs/aii-inclusion-recording-studio2021.pdf

Swim, J., Borgida, E., Maruyama, G., \& Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? Psychological Bulletin, 105(3), 409429. https://doi.org/10.1037//0033-2909.105.3.409

Ter Bogt, T. F. M., Engels, R. C. M. E., Bogers, S., \& Kloosterman, M. (2010). "Shake it baby, shake it": Media preferences, sexual attitudes and gender stereotypes among adolescents. Sex Roles: A Journal of Research, 63(11-12), 844-859. https://doi.org/10.1007/s11199-010-9815-1

Van den Berg, R. G. (n.d.-a). SPSS ANOVA - Levene's test "significant". SPSS Tutorials. Retrieved June 21, 2022, from https://www.spss-tutorials.com/spss-anova-levenes-test-significant/

Van den Berg, R. G. (n.d.-b). How to get (partial) eta squared from SPSS? SPSS Tutorials. Retrieved July 5, 2022, from https://www.spss-tutorials.com/spss-partial-eta-squared/

Watson, M., \& Anand, N. (2006). Award ceremony as an arbiter of commerce and canon in the popular music industry. Popular Music, 25(1), 41-56. https://doi.org/10.1017/S0261143005000747

Wolgast, S., Bäckström, M., \& Björklund, F. (2017). Tools for fairness: Increased structure in the selection process reduces discrimination. Plos One, 12(12), 0189512. https://doi.org/10.1371/journal.pone. 0189512

## Appendix A

Demographics
Table A.
Demographics (Extreme Outliers and Carless Responses Included)

| Characteristic | $n$ | $\%$ | $M$ | $S D$ |
| :--- | :--- | :--- | :---: | :---: |
| Gender | 123 |  | 1.77 | .42 |
| 1. Male | 28 | 22.8 |  |  |
| 2. Female | 95 | 77.2 |  |  |
| Age (in years) |  |  | 20.15 | 2.15 |
| Note. $N=123$. |  |  |  |  |

## Appendix B

## Examples of Fictitious Artist Profiles

## Male Artist Profile

Daan is a 24 -year-old male artist from Amsterdam, the Netherlands.
Please click the 'next' button to listen to a 30 -second audio snippet of one of his tracks.


Please click below to listen to the audio snippet of the track.

## Female Artist Profile

Tess is a 24 -year-old female artist from Amsterdam, the Netherlands.
Please click the 'next' button to listen to a 30-second audio snippet of one of her tracks.


Please click below to listen to the audio snippet of the track.

## Appendix C

Research Design
Table C.
Research Design

| Audio clip | Holistic group |  |  |  |  |  | Structured/Mechanical group |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | M/F |  |  | F/M |  |  | M/F |  |  | F/M |  |
|  | $n$ |  | \% | $n$ |  | \% | $n$ |  | \% | $n$ | \% |
|  | 23 |  | 24.7 | 24 |  | 27 | 18 |  | 20.2 | 25 | 28.1 |
| 1 |  | M |  |  | F |  |  | M |  |  | F |
| 2 |  | M |  |  | F |  |  | M |  |  | F |
| 3 |  | M |  |  | F |  |  | M |  |  | F |
| 4 |  | M |  |  | F |  |  | M |  |  | F |
| 5 |  | F |  |  | M |  |  | F |  |  | M |
| 6 |  | F |  |  | M |  |  | F |  |  | M |
| 7 |  | F |  |  | M |  |  | F |  |  | M |
| 8 |  | F |  |  | M |  |  | F |  |  | M |
| 9 |  | F |  |  | M |  |  | F |  |  | M |

Note. $N=89 . \mathrm{F}=$ female artist; $\mathrm{M}=$ male artist; audio clips were shown in randomized order for each participant.

## Appendix D

## Questionnaires

## Baseline

1. What is your Sona-identification number?
2. What is your age?
3. What gender do you most identify with?
1) Male
2) Female
3) Other/Prefer not to say

## Table D1.

Items for Music Assessment

| Experimental group | Item |
| :---: | :---: |
| Holistic | 1. How much did you like the track you |
|  | just listened to? ${ }^{\text {a }}$ |
| Structured/mechanical | 1. I think the track is rhythmical. ${ }^{\text {b }}$ |
|  | 2. I think the track conveys a nice |
|  | mood. ${ }^{\text {b }}$ |
|  | 3. I think the track is innovative. ${ }^{\text {b }}$ |
|  | 4. I think the track is interesting. ${ }^{\text {b }}$ |
|  | 5. I think the track is expressive. ${ }^{\text {b }}$ |
|  | 6. I think the track reflects technical |
|  | competence. ${ }^{\text {b }}$ |
|  | 7. I think the track conveys artistic |
|  | merit. ${ }^{\text {b }}$ |

Table D1 Continued.

| Experimental group | Item |
| :---: | :---: |
| Structured/mechanical | 8. How much did you like the track you just listened to? ${ }^{\text {a }}$ |
| Note. |  |
| ${ }^{\text {a }}$ Ratings were on 7-point Likert scales ranging from $1=$ dislike a great deal to $7=$ like a |  |
| ${ }^{\mathrm{b}}$ Ratings were on 7-po agree. | from $1=$ strongly disagree to $7=$ strongly |

Table D2.
Items Assessing Gender Stereotypes Against Male and Female Music Artists

| Dimension | Item |
| :--- | :--- |
| Rhythm | Music by male artists is usually more |
| Mood | rhythmical than music by female artists. |
|  | Music by male artists usually expresses a |
|  | song's mood better than music by female |
| Innovativeness | artists. |
|  | Music by female artists is usually more |
| Interestingness | innovative than music by male artists. |
|  | Music by male artists is usually more |
|  | interesting than music by female artists. |

Table D2 Continued.

| Dimension | Item |
| :--- | :--- |
| Artistic Merit | Music by female artists usually conveys |
|  | more artistic merit than music by male |
|  | artists. |
| Expressiveness | Music by female artists is usually more |
|  | expressive than music by male artists. |
| Technical Competence | Music by female artists usually reflects less |
|  | technical competence than music by male |
|  | artists. |

Note. Ratings were on a 7-point Likert scale ranging from $1=$ strongly disagree to $7=$ strongly agree.

Appendix E<br>Means and Standard Deviations

## Table E1.

Mean Music Ratings and Standard Deviations for Analyses of Hypotheses 2 and 3
(Excluding Extreme Outliers and Careless Responses)


Table E1 Continued.

| Artist gender | Assessment approach | Participant gender | $n$ | M | $S D$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Male |  | Female | 30 | 4.51 | . 7 |
|  | Mechanical ${ }^{\text {b }}$ | Male | 13 | 4.4 | . 75 |
|  |  | Total | 43 | 4.48 | . 71 |
|  | Total | Male | 21 | 4.15 | . 86 |
|  | (holistic and | Female | 69 | 4.45 | . 73 |
|  | mechanical) | Total | 89 | 4.38 | . 77 |
|  | Structured ${ }^{\text {a }}$ | Total | 43 | 3.98 | 1.02 |
|  | Total | Total | 89 | 4.14 | . 93 |
|  | (holistic and structured) |  |  |  |  |

Note. $N=89$. Means for mechanical and structured assessment were computed for the same subgroup sample.
${ }^{\text {a }}$ Mean music ratings were based on one item for each track measuring liking on a 7-point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal .
${ }^{\mathrm{b}}$ Mean music ratings were based on the average score of ratings for the seven assessment criteria for each track, each measured on a 7-point Likert scale ranging from $1=$ strongly disagree to $7=$ strongly agree; The final mean music rating score is measured on a 7 -point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal.

Table E2.
Mean Music Ratings and Standard Deviations for Analyses of Hypotheses 2 and 3
(Including Extreme Outliers and Careless Responses)


Table E2 Continued.

| Artist gender | Assessment <br> approach | Participant gender | $n$ | M | SD |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Male | Mechanical ${ }^{\text {b }}$ | Total | 59 | 4.53 | . 86 |
|  | Total | Male | 28 | 4.0 | 1.11 |
|  | (holistic and | Female | 95 | 4.48 | . 89 |
|  | mechanical) | Total | 123 | 4.38 | . 96 |
|  | Structured ${ }^{\text {a }}$ | Total | 59 | 4.07 | 1.14 |
|  | Total | Total | 123 | 4.16 | 1.09 |
|  | (holistic and |  |  |  |  |
|  | structured) |  |  |  |  |

Note. $N=123$. Means for mechanical and structured assessment were computed for the same subgroup sample.
${ }^{\text {a }}$ Mean music ratings were based on one item for each track measuring liking on a 7 -point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal.
${ }^{\mathrm{b}}$ Mean music ratings were based on the average score of ratings for the seven assessment criteria for each track, each measured on a 7-point Likert scale ranging from $1=$ strongly disagree to $7=$ strongly agree; The final mean music rating score is measured on a 7 -point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal.

Table E3.
Means and Standard Deviations of Music Ratings for Repeated Tracks (Excluding
Extreme Outliers and Careless Responses)

| Artist gender | Assessment <br> approach | Participant gender | $n$ | M | $S D$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Female |  | Male | 8 | 4.0 | . 93 |
|  | Holistic ${ }^{\text {a }}$ | Female | 38 | 4.71 | 1.21 |
|  |  | Total | 46 | 4.59 | 1.19 |
|  |  | Male | 13 | 4.62 | . 73 |
|  | Mechanical ${ }^{\text {b }}$ | Female | 30 | 4.73 | . 81 |
|  |  | Total | 43 | 4.69 | . 78 |
|  | Total | Male | 21 | 4.38 | . 85 |
|  | (holistic and | Female | 68 | 4.72 | 1.04 |
|  | mechanical) | Total | 89 | 4.64 | 1.01 |
|  | Structured ${ }^{\text {a }}$ | Total | 43 | 4.42 | 1.48 |
|  |  | Total | 89 | 4.51 | 1.33 |
|  | (holistic and |  |  |  |  |
|  | structured) |  |  |  |  |
|  |  | Male | 8 | 3.88 | 1.13 |
| Male | Holistic ${ }^{\text {a }}$ | Female | 38 | 4.5 | 1.13 |
|  |  | Total | 46 | 4.39 | 1.15 |
|  |  | Male | 13 | 4.45 | . 98 |
|  | Mechanical ${ }^{\text {b }}$ | Female | 30 | 4.59 | . 85 |
|  |  | Total | 43 | 4.54 | . 88 |

Table E3 Continued.

| Artist gender | Assessment approach | Participant gender | $n$ | M | $S D$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Male | Total | Male | 21 | 4.23 | 1.05 |
|  | (holistic and | Female | 68 | 4.54 | 1.01 |
|  | mechanical) | Total | 89 | 4.47 | 1.02 |
|  | Structured ${ }^{\text {a }}$ | Total | 43 | 4.28 | 1.59 |
|  | Total | Total | 89 | 4.34 | 1.37 |
|  | (holistic and |  |  |  |  |
|  | structured) |  |  |  |  |

Note. $N=89$. Means for mechanical and structured assessment were computed for the same subgroup sample.
${ }^{\text {a }}$ Mean music ratings were based on one item for each track measuring liking on a 7-point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal .
${ }^{\mathrm{b}}$ Mean music ratings were based on the average score of ratings for the seven assessment criteria for each track, each measured on a 7-point Likert scale ranging from $1=$ strongly disagree to $7=$ strongly agree; The final mean music rating score is measured on a 7 point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal.

Table E4.
Means and Standard Deviations of Music Ratings for Repeated Tracks (Including Extreme Outliers and Careless Responses)

| Artist gender | Assessment approach | Participant gender | $n$ | M | $S D$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Female |  | Male | 13 | 4.08 | 1.44 |
|  | Holistic ${ }^{\text {a }}$ | Female | 51 | 4.8 | 1.28 |
|  |  | Total | 64 | 4.66 | 1.34 |
|  |  | Male | 15 | 4.57 | . 84 |
|  | Mechanical ${ }^{\text {b }}$ | Female | 44 | 4.78 | 1.05 |
|  |  | Total | 59 | 4.73 | 1.0 |
|  | Total | Male | 28 | 4.34 | 1.16 |
|  | (holistic and | Female | 95 | 4.79 | 1.18 |
|  | mechanical) | Total | 123 | 4.69 | 1.18 |
|  | Structured ${ }^{\text {a }}$ | Total | 59 | 4.51 | 1.56 |
|  | Total | Total | 123 | 4.59 | 1.44 |
|  | (holistic and structured) |  |  |  |  |


|  | Male | 13 | 4.0 | 1.73 |
| :---: | :---: | :---: | :---: | :---: |
| Male | Holistic $^{\text {a }}$ | Female | 51 | 4.57 |
|  |  | Total | 64 | 4.45 |
|  |  | Male | 15 | 4.47 |
|  |  | Mechanical ${ }^{\text {b }}$ | Female | 44 |
|  | Total | 59 | 4.65 | 1.43 |
|  |  |  | 4.6 | 1.06 |

Table E4 Continued.

| Artist gender | Assessment | Participant | $n$ | $M$ | $S D$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  | approach | gender |  |  |  |
|  | Total | Male | 28 | 4.25 | 1.35 |
|  | (holistic and | Female | 95 | 4.61 | 1.21 |
|  | mechanical) | Total | 123 | 4.52 | 1.25 |
|  | Structured ${ }^{\text {a }}$ | Total | 59 | 4.36 | 1.65 |
|  | Total | Total | 123 | 4.41 | 1.53 |
|  | (holistic and |  |  |  |  |
|  | structured) |  |  |  |  |

Note. $N=123$. Means for mechanical and structured assessment were computed for the same subgroup sample.
${ }^{\text {a }}$ Mean music ratings were based on one item for each track measuring liking on a 7 -point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal .
${ }^{\mathrm{b}}$ Mean music ratings were based on the average score of ratings for the seven assessment criteria for each track, each measured on a 7-point Likert scale ranging from $1=$ strongly disagree to $7=$ strongly agree; The final mean music rating score is measured on a 7-point

Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal.

## Appendix $F$

Results of Exploratory Analysis for Effect of Participant Gender on Gender Bias

## Figure F.

Mean Ratings by Female and Male Participants


Note. $N=89$. Holistic condition: $n=46$, structured/mechanical condition: $n=43$, male participants: $n=21$, female participants: $n=68$, mean music rating scores are on a 7 -point Likert scale ranging from $1=$ dislike a great deal to $7=$ like a great deal, $\mathrm{F}=$ female artist, M $=$ male artist, ratings were on 7-point Likert scales ranging from $1=$ dislike a great deal to 7 $=$ like a great deal, error bars: +/- 2 SD.

## Appendix G

Analyses for Sample Including Extreme Outliers and Careless Responses

## Table G1.

Analysis of Anti-Female Stereotypes (Hypothesis 1)

|  | $M$ | $S D$ | $t(122)$ | $p$ | $d$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Stereotype   <br> ratings 3.61 .5 <br>   80.65 |  |  | .00 | .78 |  |

Note. $N=123$. The mean score difference from the scale midpoint (4) is tested. $p<.05$.

Table G2.
Analysis of Effect of Participant Gender on Anti-Female Stereotypes (Exploratory
Hypothesis 1)


Note. $N=123$. There were 28 male participants and 95 female participants in the sample. $p<.05$.

## Table G3.

Analysis of Effects of Artist Gender and Assessment Approach (Holistic and Mechanical) on Music Ratings (Hypothesis 2 and 3)

|  | $S S$ | $d f$ | $M S$ | $F$ | $p$ | $\eta_{p}{ }^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Main effects |  |  |  |  |  |  |
| Artist gender | 6.29 | 1 | 6.29 | 3.8 | .05 | .03 |
| Assessment |  |  |  |  |  |  |
| Interaction effect | .04 | 1 | .04 | .23 | .63 | .00 |

Note. $N=$ 123. $S S=$ Type III Sum of Squares; $M S=$ Mean Square. $p<.05$.

## Table G4.

Analysis of Effects of Participant Gender, Artist Gender, and Assessment Approach
(Holistic and Mechanical) on Music Ratings (Exploratory Hypothesis 2 and 3)

|  | $S S$ | $d f$ | $M S$ | $F$ | $p$ | $\eta_{p}{ }^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Main effects |  |  |  |  |  |  |
| Artist gender | 8.58 | 1 | 8.58 | 5.3 | .02 | .04 |
| Assessment | 6.37 | 1 | 6.37 | 3.93 | .05 | .03 |
| Participant gender |  |  |  |  |  |  |
| Interaction effects | .19 | 1 | .19 | 1.14 | .29 | .01 |
| Artist gender * assessment | .77 | 1 | .77 | 4.57 | .04 | .04 |
| Artist gender * participant |  |  |  |  |  |  |
| gender | 1.08 | 1 | 1.08 | 6.44 | .01 | .05 |
| Artist gender * assessment * |  |  |  |  |  |  |
| participant gender |  |  |  |  |  |  |

Note. $N=$ 123. $S S=$ Type III Sum of Squares; $M S=$ Mean Square. There were 28 male participants and 95 female participants in the sample.
$p<.05$.

## Table G5.

Analysis of Effects of Artist Gender and Assessment Approach (Holistic and Mechanical) on Music Ratings for Repeated Audio Clips (Hypothesis 2 and 3)

| $S S$ | $d f$ | $M S$ | $F$ | $p$ | $\eta_{p}{ }^{2}$ |
| :--- | :--- | :--- | :--- | :--- | :--- |

Main effects

| Artist gender | 1.64 | 1 | 1.64 | 4.36 | .04 | .04 |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Assessment | .74 | 1 | .74 | .29 | .59 | .00 |

Interaction effect
Artist gender * assessment 11 1 11 . 26 . 61 . 00
Note. $N=$ 123. $S S=$ Type III Sum of Squares; $M S=$ Mean Square . $p<.05$.

## Table G6.

Analysis of Effects of Artist Gender and Assessment Approach (Holistic and Structured) on Music Ratings (Exploratory Hypothesis 3)

|  | $S S$ | $d f$ | $M S$ | $F$ | $p$ | $\eta_{p}{ }^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Main effects |  |  |  |  |  |  |
| Artist gender | .45 | 1 | .45 | 1.53 | .22 | .01 |
| Assessment | .77 | 1 | .77 | .39 | .53 | .00 |
| Interaction effect |  |  |  |  |  |  |
| $\quad$ Artist gender * assessment | .16 | 1 | .16 | .53 | .47 | .00 |

Note. $N=$ 123. $S S=$ Type III Sum of Squares; $M S=$ Mean Square. $p<.05$.

## Table G7.

Analysis of Effects of Artist Gender and Assessment Approach (Holistic and Structured) on Music Ratings for Repeated Audio Clips (Exploratory Hypothesis 3)

|  | $S S$ | $d f$ | $M S$ | $F$ | $p$ | $\eta_{p}{ }^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Main effects |  |  |  |  |  |  |
| Artist gender | .92 | 1 | 1.94 | 3.05 | .08 | .03 |
| Assessment |  |  | .92 | .24 | .62 | .00 |
| Interaction effect | .04 | 1 | .04 | .06 | .8 | .00 |
| $\quad$ Artist gender * assessment |  |  |  |  |  |  |

Note. $N=$ 123. $S S=$ Type III Sum of Squares; $M S=$ Mean Square. $p<.05$.

## Table G8.

Analysis of Effects of Participant Gender, Artist Gender, and Assessment Approach (Holistic and Mechanical) on Music Ratings for Repeated Audio Clips (Exploratory Hypothesis 2 and 3)

|  | $S S$ | $d f$ | $M S$ | $F$ | $p$ | $\eta_{p}{ }^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Main effects |  |  |  |  |  |  |
| Artist gender | .8 | 1 | .8 | 2.11 | .15 | .02 |
| Assessment | 2.78 | 1 | 2.78 | 1.09 | .3 | .01 |
| Participant gender | 7.65 | 1 | 7.65 | 2.99 | .09 | .03 |

Interaction effects

| Artist gender * assessment | .02 | 1 | .02 | .04 | .84 | .00 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Artist gender * participant | .09 | 1 | .09 | .24 | .63 | .00 |
| gender |  |  |  |  |  |  |
| Artist gender * assessment * | .05 | 1 | .05 | .13 | .72 | .00 |
| participant gender |  |  |  |  |  |  |

Note. $N=$ 123. $S S=$ Type III Sum of Squares; $M S=$ Mean Square. There were 28 male participants and 95 female participants in the sample. $p<.05$.


[^0]:    ${ }^{1}$ For this criterion, one track presented as made by a female artist in one of the two mechanical conditions was by error rated on a five-point Likert scale. Therefore, this item was recoded to a seven-point Likert scale but cannot contain the values for $2=$ disagree, and $5=$ agree. This is not expected to distort the results extensively.

