Gender Bias in Competitive Music Composition Evaluation: An Experimental Study

Chaowen Ting, Georgia Institute of Technology
Yating Chuang, National Taipei University
John Chung-En Liu, National Taiwan University

May 10, 2020

Abstract

Women's underrepresentation in science has drawn attention from scholars in many disciplines. Our study aims to advance understanding on this issue through examining the extreme case of gender imbalance in music composition—a field in which PhDs have been awarded to women at a rate lower than that in computer science, physics, economics, and philosophy. To examine whether such gender imbalance is due to unfair judgment in evaluative settings, we conducted a unique experiment in which newly commissioned orchestral compositions randomly attributed to female and male composers were evaluated. We invited college composition faculty to rate the compositions based on the music scores along with live recordings directed by the same conductor to eliminate pre-exiting bias. Contrary to our hypotheses, compositions associated with female names were rated higher than compositions associated with male names. We also find strong evidence of gender out-group bias among male reviewers. When we further analyze how the results vary across faculty demographics, we find that male faculty, faculty over age 45, assistant professors, and full professors favored female composers in both general and structured evaluations, while female faculty, faculty aged below 45, adjunct faculty and associate professors did not show significant preference toward either gender.

1. Introduction

The persistent underrepresentation of women in science has raised concern in academic communities around the world. In the United States, women account for 46.6% of all earned doctorate recipients, 45% of associate professors, and on 33% of full professors in 2017.[1] This issue, often highlighted in STEM (science, technology, engineering, and math) fields, is also of great concern in non-STEM field. For example, in economics, under 35 percent of PhD students and 30 percent of assistant professors are female by the mid-2000s (Lundberg and Stearns 2019). Such gender imbalance has stimulated lively discussion and new research, to the extent that the *Journal of Economic Perspective* hosted a special symposium on women in economics in 2019 (Bayer and Rouse 2016; Avilova and Goldin 2018; Boustan and Langan 2019; Lundberg and Stearns 2019; Buckles 2019). One leading explanation of the gender imbalance is that female candidates are not equally evaluated in settings such as interviews and peer-reviews. Such

[1] The National Science Foundation and National Center for Education Statistics. https://nces.ed.gov/programs/coe/indicator_csc.asp

discrimination can be caused by biases against minority groups (Greenwald and Banaji 1995; Greenwald and Krieger 2006; Bayer and Rouse (2016)[2]) or "in-group bias"—people favor in-group members, such as men favoring other men in a male-dominated field (Tajfel et al., 1971; Bernhard et al., 2006; Chen and Li, 2009; Goette et al., 2012; Sandberg 2017). Either way, women face disadvantages and are less likely to advance in their careers.

We believe that a study of professional music composition may offer new insights on the origins of such gender representation disparity. A field not requiring intensive math skills, severe gender disparity exists in the profession of music composition: in 2011, only about 16% of U.S. doctorates were awarded to women. This number is the lowest in women's representation among all studied disciplines, including physics (18%), computer science (19%), and economics (35%) (Leslie et al. 2015: Fig. 1). This gender gap grows even wider in the professional realm. Women held only 15% percent of the U.S. composition faculty positions in 2014, and only 9% of the prestigious composition prizes had been awarded to women (Q2 Music 2014). A recent study revealed that only 2% of the works performed by symphony orchestras worldwide in 2018 – 2019 were written by women (Brown 2018).

We designed an audit experiment examining whether college composition faculty members evaluate compositions fairly when the music is randomly associated with a gendered name, and whether more structured evaluation methods, such as giving detailed ratings on various musical dimensions, could reduce the bias. We recruited composition faculty members at U.S. higher education institutions to evaluate four orchestral compositions as if they were judging a composition competition. Faculty members were recruited from 377 music institutions that participated in the 2013-14 National Association of Schools of Music (NASAM) Higher Education Arts Data Services (HEADS) project, and all of the participants have a training background in composition. Participants were randomly assigned to one of the two groups, and the randomization was done at the institutional level, meaning that the faculty at the same school were assigned to the same survey. In Survey Group A, compositions appeared with the gendered name sequence of M, F, M, F. In Survey Group B, the exact same compositions appeared with the opposite sequence of F, M, F, M. All music pieces are brand new, last between 8 to 10 minutes, and were composed by four different professional composers with a doctoral degree in composition. In the evaluation, we provided both the music scores and the audio recordings from live performances under the same conductor and university orchestras to ensure the quality of recordings and performances were equal. Our experimental design replicates the setting of similar competitions.

This paper has several contributions to the literature. First, we contribute to the literature on discrimination. A substantial amount of studies in this literature finds evidence of discrimination against racial, ethnic, and gender minority groups (see the review of field experiments on discrimination by Betrand and Duflo 2016). The leading economic models to explain discrimination are 1) Becker's taste-based model (Becker, 1957; Charles and Guryan, 2008) and 2) statistical discrimination model (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977). In Becker's model in the labor market context, some people have non-pecuniary distaste of hiring members from the minority groups. The statistical discrimination model is based on the rational

[2] Bayer and Rouse (2016) provides a review in this demand side explanation for understanding diversity in the economics profession.

perception of skill distribution between different groups, and predicts employers/evaluators would favor the majority groups (e.g. white and male) because of their perceived higher productivity. In both models, we may see gender bias favoring male, as documented in the literature on underrepresentation of women in academia (for surveys of the evidence: see Greenwald and Banaji 1995; Greenwald and Krieger 2006; Bayer and Rouse 2016). However, a large number of the field experiments lack the information of employers' identities, making it is difficult to examine the mechanism to explain the sources of discrimination.

To understand the specific mechanisms of discrimination, *in-group bias* has become another prominent explanation. Our experiment, one that mimics a real-world music competition, can add to this literature. Most of the evidence on in-group bias was done in the laboratory setting. Laboratory experiments have shown that in-group favoritism will arise in both natural social groupings or artificial created groupings (Tajfel et al., 1971; Bernhard et al. 2006; Leider et al., 2009; Chen and Li, 2009; Chen and Chen, 2011; Goette et al., 2012; Chen et al, 2014; Currarini and Mengel, 2016; Dickinson et al. 2018). It is harder, however, to test whether this favoritism would hold in a real-world setting. So far, studies have found mixed results. For example, in the professional sports setting, scholars have found discrimination toward the other racial ethnic groups (Price and wolfers 2010; Parsons et al. 2011), and favoritism to the same nationality, but no in-group bias in favor of gender (Sandberg 2017). In judicial decision-making, there is evidence of racial or ethnic in-group bias (Shayo and Zussman, 2011, 2017; Anwar *et al*., 2012), while Depew et al (2017) find opposite result of racial out-group bias using juvenile court cases. Similar racial in-group bias can also be found in policing (Antonovics and Knight, 2009; West, 2018), lending decisions (Fisan et al. 2017), and equity analysis (Jannati et al. 2018), while some research suggests no gender favoritism in journal reviews (Abrevaya and Hamermesh 2012) and grading (Jeld et al. 2015), even out-group bias in Bar and Zussman's (2020) study of driving tests. In sum, the existence and extent of in-group bias on gender, race, and nationality is far from conclusive and demands more empirics.

Our experiment crafts a rather "clean" non-laboratory setting to test in-group bias. We will outline below why it is challenging to test in-group bias outside of the laboratory setting, and how our design can circumvent these issues.

First, the evaluator's identity is often unknown in the real-world setting. In most of the experimental studies in testing discrimination, researchers can only collect the outcome such as the callback rate in Bertrand et al. (2005) (See more review on field experiments in testing discrimination by Bertrand and Duflo 2016). In Goldin and Rouse's (2000) examination of audition outcomes, they had the similar problem of unavailable information on the juries' identities. Thus, it is impossible to test the source of discrimination. In our study, we recruited the faculty reviewers based on certain criteria to make sure we select the appropriate sample of reviewers. We have various male and female reviewers and the demographics are similar in the two experimental groups.

Second, there could be a correlation between evaluators' associated group identity (e.g. gender) and candidates' quality. In other words, the matching between evaluators and candidates is not random, and thus hard to verify the existence of in-group bias. This can be a concern especially in observational studies, for example, in the context of journal review (Abrevaya and

Hamermesh 2012), tenure review (Bagues and Esteve-Volart, 2010), or professional sports competition (Price and Wolfers 2010; Sandberg 2017). In our audit experiment, all faculty reviewers were assigned to four identical compositions with randomly assigned gendered names (two female and two male names). This matched gender identity between evaluators and candidates can circumvent those identification challenges.

Third, the group identity (race and gender) is not exogenous, and thus the discriminatory effects in the market outcome could be overstated due to other unobservable variables (Siegelman and Heckman 1992; Heckman 1998). In other words, it is hard to examine the causal effect of being "female" or "white" while keeping all else constant. In most non-laboratory studies, for example, the actual productivity/skill of the candidate is rarely observed directly and can confound with the minority identities. In our experiment, we can circumvent this problem since the same composition is always associated with both female and male names. For example, composition #1 is assigned with a male name in randomization group A and with a female name in group B. The unobservable characteristics is averaged out to zero for the same audit pair.

Fourth, another confound can come from endogenous behaviors in typical hiring and evaluation studies. For example, in hiring studies, it is possible that the interviewees' behaviors may be endogenous to recruiter's race or gender. Even in audit studies, the trained interviewees know the purpose of the study and may thus unconsciously want to do a "good job." This demand side effect from behavioral change can generate large confounding factor to the discriminatory effect. In our field experiment, we only rely on music scores and recordings rather than actual people, and thus are able to eliminate this endogenous behavioral factor and truly distinguish gender bias or in-group bias.

Our result shows evidence of an *out-group* bias by male evaluators in favor of female composers in the music composition. In addition, more experienced professors (tenured professors and older professors) show preference toward female composers than less experienced professors (untenured professors and younger professors). This pattern also persists in the more structured evaluations when participants assess compositions in different subcategories. Our results align with various studies that fail to find gender favoritism in various contexts such as academic evaluations (Broder 1993; Bagues et al. 2017), hiring in the Spanish Judiciary (Bagues and Esteve-Volart 2010), female dominated occupations in Australia (Booth and Leigh 2010),[3] and driving test in Israel (Bar and Zussman 2020).

There are several implications. First, biases in treating the minority group may be an overstated explanation for the underrepresentation of women in some highly gender imbalanced fields. We do not find this gender stereotype pattern suggested by the "statistical discrimination model" that minority candidates (i.e. female composers) should be treated less favorably due to imperfect information and expected productivity. Second, in-group bias hypothesis may not hold outside laboratories as this preference may differ in contexts. Third, in a highly professional setting where gender imbalance persists, people may evolve preference toward diversity.

[3] One has to be careful interpreting the result in Booth and Leigh (2010) as opposite gender favoritism because the hiring experiment is conducted in female dominated jobs, such as waitstaff, data-entry, customer service, and sales.

The next section describes the contexts of composition competitions. Section 3 and 4 introduce the experimental design, and the estimation strategy. Section 5 discusses the results. Section 6 presents the robustness checks. Section 7 concludes.

2. Music Composition Competition

We focused on composition competitions because they are important for composers' career advancement. Such events are the primary vehicle to identify young talents ready to launch international careers (McCormick 2009). The competitions are frequently hosted by various organizations and ensembles: a quick visit to *The Composer's Site*, an online information hub for composers, yielded 63 ongoing competitions in June 2020. A "call for scores" or composition competition is similar to an academic conference where young professionals present their works, build professional networks, establish creative portfolios, and engage with potential employment opportunities. Composers' works are selected to be performed through such competitions, and performances of these kinds often leads to future performances, engagements, or even commissions (Whitacre 2009; Watts 2018; Murphy 2012; Doolittle 2018). In addition, evidence has shown that the prior success in the music competitions has significant impact on the later success. Though in a different context, Victor et al. (2003)'s study on the participants of the Queen Elizabeth Competition, one of the most prestigious piano and violin competitions, showed that participants' prior success due to random factors lead to subsequent market success.

Our experimental design follows the form of hosting a formal composition competition. The organizer institution send out invitations to invite professional composers as reviewers. The scores, often along a recording, will be uploaded in a platform and the reviewers will give evaluations of the compositions. In theory, the composer's identity can be anonymous in the competition, but in practice, it is challenging to remain anonymity in this tightly-connected field. Various esteemed and highly influential composition competitions, such as Guggenheim, Fromm, Koussevitzky, ASCAP, Barlow, are not anonymous, which we follow suit. Although professional composers can judge the quality of the composition simply based on the music score, it is common for the competition to provide the recordings to facilitate the evaluation process. We adopted the same design in this paper. One of the authors, a professional conductor, conducted and recorded all four music pieces to ensure the quality of recordings are comparable to professional standards.

3. Experimental Design

In this section, we present the experimental design to measure gender bias in the professional music composition setting. We also aim to test whether there is same-gender in-group bias as found in the experimental literature. As some research suggest that structured evaluation methods—ones that with clear rubrics—could reduce bias, we also test this claim in the experiment (Bragger et al. 2002; Brecher et al. 2006; Levashina et al. 2014). Our hypotheses were:

Hypothesis A: Reviewers favor compositions associated with a male name (gender discrimination hypothesis).

Hypothesis B: Reviewers would show gender bias toward the same gender (favoritism/in-group bias hypothesis).

Hypothesis C: Evaluation discrimination against female composers would be reduced or eliminated by structured evaluation.

Constructing the audit experiment in composition

Our main objective is to detect gender bias in music competitions by using compositions randomly assigned to male and female composer names. All compositions last between 8 to 10 minutes long, and were composed by four different professional composers with a terminal doctoral degree in composition. All of the works were chosen by the same conductor using the identical selection criteria, and they were all commissioned works by the Georgia Tech School of Music. As they have only been performed once in a regional concert in southwestern U.S., reviewers were agnostic about the compositions before our experiment. This is common practice in competitions that composers submit recordings of performance or reading sessions at their affiliated institutions for consideration. Consistent with real composition competitions, we provided both the score and the recording of each piece for reviewers rating. All the recordings were made during two performances in the same concert hall conducted by the same conductor within two months. As male composers are more dominated and established in this field, the recording quality and orchestral performance may sometimes endogenously correlate with this gender factor, so this design makes sure that we isolate these other noisy factors.

To test gender bias, we assigned two female (F) names and two male names (M) to the four compositions. The designated composers' names only appeared on the title page and the first page of the music score as the standard practice in composition competitions. We pre-tested the survey with faculty members that have served as composition competition judges. They agreed that this design was similar to real composition competitions and none detected that gender bias was being tested. With limited gender cues, respondents in our pre-tested survey were able to correctly recall the gender of the composers, indicating that subjects were truly "treated" in this design. A sample page of the score is in the Appendix.

Names were selected from the U.S. Social Security Administration's database of popular names in the 1970s with the most common last names from the 2000 census. We pre-tested and google searched the names to make sure that 1) gender cue is easily recognized; 2) last names do not give extra racial/ethnic cues; 3) no actual composers have the same names in our experiment. We alternated the female and male names in two groups, A and B. The name assignment in group A is M, F, M, F, namely Michael Adams, Rebecca Moore, Sean Campbell, Tara Davis. Names used in group B were F, M, F, M, namely Tara Davis, Sean Campbell, Rebecca Moore, Michael Adams.

Recruitment

Faculty members were recruited from 377 music institutions that participated in the 2013-14 National Association of Schools of Music (NASAM) Higher Education Arts Data Services (HEADS) project. To identify qualified reviewers, we searched the websites of these music

schools and reviewed faculty members' profiles and CVs to identify all eligible participants with composition expertise, either composing experience or a music composition degree. We included faculty of all tenure-track ranks (assistant professors, associate professors, and full professors) as well as adjunct faculty. The final invitation list contained 1,060 eligible faculty members. Faculty reviewers were randomly assigned to group A or B to evaluate four compositions as if they were evaluating a composition competition. Randomization was at the institution level to ensure that faculty members at any institution received compositions associated with the same artificial female and male composer names to prevent sample contamination. We also cross-checked for duplicate faculty names in the case that the same person is in transition from one institution to another, or that certain faculty is affiliated with more than two institutions to ensure that every faculty is only invited once under one randomization group.

The survey was conducted between March and May 2018. Eligible faculty were invited to voluntarily participate (see the invitation email in Appendix: Survey Materials for reference) via Qualtric system through a personalized survey link in the invitation email. They were rewarded a $75 gift card upon completion of the evaluation. Only the invited eligible faculty were able to complete an evaluation and receive a gift card for completing the survey. This personalized link ensures the highest data quality and prevents sample corruption from ineligible faculty.

The final sample contained 124 faculty reviewers (20 female, 103 male, 1 other). The response rate is 11.7%, which is not uncommon for a highly selected group. Literature has also shown that the result of the focal variable is not sensitive to the response rate if the sample demographics is similar to the underlying population (Williams and Ceci 2015). Also, since we use the same procedure recruiting faculty reviewers in two randomized groups (the recruiting emails were sent on the same day with the same content), there is no differential difference in the attrition rate. The summary statistics of the final sample is in Table 1. We tested the difference in means among various covariates between randomization group A and B. The covariates include demographic variables, such as gender, race, and, rank. We have also tested the average overall evaluation score between group A and B. As Table 1 shows, our randomization works well since none of the differences between two groups appear to be statistically significant.

Evaluation Survey

We disguised this survey as a simple music evaluation study without revealing to reviewers that this was a gender bias experiment. Reviewers were asked to judge the compositions as if they were judging a real competition. The composer names only appeared in the score and the final question where reviewers were asked to recommend a winner. This avoids experimenter effect that people may behave very differently when they know that they were in an experiment testing gender bias.

The survey contained four parts: general evaluation, structured evaluation, winner recommendation, and demographic information. In conventional practice, overall evaluation is often the only judgement the reviewers give. Thus, the interpretation of gender bias should be highly weighed on this score. Overall evaluation was scored on a 0-10 scale.

Overall evaluation can be highly subjective. Therefore, we added a structured evaluation in Part II to evaluate compositions in more details. Presumably, this structured evaluation would prompt reviewers considering the works more carefully when judging compositions and may eliminate arbitrary bias, if any. Part II evaluation consisted of five categories, namely (1) Form / Structure, (2) Tonality / Harmony, (3) Tempo / Rhythm, (4) Orchestration, and (5) Artistic Originality. These five dimensions were chosen as they consist of the basic elements of musical composition, and were suggested by professional composers who have judged composition competitions. Those five categories covered both crafts and creativity of music composition, and were gender neutral terms. To ensure the integrity of the Part II evaluation, reviewers were given the score and recordings again and were not allowed (by our pre-set Qualtrics algorithm) to change their initial overall scores in Part I once they reached Part II.[4] The structured evaluation in Part II were on a 1-5 Likert scale labeled "Extremely Weak," "Somewhat Weak," "Neutral," "Somewhat Strong," and "Extremely Strong." The scale was pretested among faculty who have judged music and they agreed that this chosen scale is standard. The survey questions are in Appendix: Survey Materials.

4. Estimation Strategy

In the summary statistics and the mean evaluation tests, we used the original 124 faculty reviewers as the unit of analysis to show the overall descriptive characteristics in the data. To perform gender bias analysis, we treated the unique combination of composition and reviewer as the unit of analysis. That said, the total observations expanded into N=496. As each reviewer evaluated four compositions and the standard errors may be correlated within the same reviewers, we followed the standard practice of clustering our standard errors by reviewer, allowing more conservative (larger) standard errors to perform our statistical test. Our estimation equation is:

$$Score_{ic} = \alpha + \beta Female_{ic} + \varepsilon_{ic}$$

where $Score_{ic}$ is the evaluation score by reviewer i for composition c; $Female_{ic}$ is the dummy variable where it is equal to 1 if the composition is randomly assigned with a female name, and 0 if the composition is assigned with a male name; α is the constant term; β is the coefficient for gender bias. We expected $\beta < 0$ if there is implicit gender bias in favor of male composers, relative to female composers. We ran similar analysis using 7 different evaluation scores as the dependent variable, namely, the overall score in survey part I, the structured scores in 5 categories in survey part II, and the average score of the 5 categories in part II. We also estimated those similar 7 sets of regressions by gender, age, and rank to examine heterogeneity in gender bias. In the robustness check section, β is also estimated under different fixed effects models.

[4] The other potential experiment design is to separate Part II evaluation from Part I into other two randomization groups. Since the common evaluation practice always includes Part I-the overall evaluation, this design will make the survey too far from the composition competition norm.

To examine in-group bias, we estimate whether those reviewers would give higher scores to the compositions assigned with names same as their own gender, as opposed to compositions assigned with opposite gendered names. Similar to the above equation, we estimate:

$$Score_{ic} = a + bSame\_gender_{ic} + e_{ic}$$

where $Same\_gender_{ic}$ is the dummy variable which is equal to 1 if the assigned composers' names and the reviewers are of the same gender, and 0 if otherwise. If there is in-group bias among people within the same membership (i.e. gender), b should be larger than zero. We also controlled for several fixed effects in the robustness check section.

5. Results

Based on the ratings of the same compositions with male and female names, we estimated gender bias through the standard regression analysis clustering standard errors. We also employed further control for variables that we worried might bias our results. Below we describe detailed results of the analysis.

A. Is there gender bias in ratings?

Contrary to hypothesis A, female composers were favored in all evaluations. Compositions appearing with female names received higher scores in both of the general rating and structured rating. In overall ratings, female composers scored on average 0.294 points higher on a 0-10 scale (See Fig. 1 and Table 2). This differential bias is non-negligible—about 15.3% of the overall standard deviation. For example, in another professional evaluation context, Dressage in Olympics, the size of own-nationality bias is between 7.2% and 23.8% of the overall standard deviation (Sandberg 2017). In addition, when we ask the reviewers to pick the winners, the compositions with female names were 1.6 times more likely to be selected as the winners, compared to the same compositions associated with male names.

In the structured evaluation, differences were found in all five categories, with female composers receiving an average of 0.182 more points on a 1-5 scale (See Fig 2 and Table 2). Specifically, the difference was 0.153 in Form / Structure, 0.197 in Tonality / Harmony, 0.073 in Tempo / Rhythm, 0.242 in Orchestration, and 0.246 in Artistic Originality (all are statistically significant except for Tempo / Rhythm; See Fig.2 and Table 2). These differential numbers favoring female composers correspond to 14.7% of the overall standard deviation in Form / Structure, 18.6% of the overall standard deviation in Tonality / Harmony, 6.8% of the overall standard deviation in Tempo / Rhythm, 21.4% of the overall standard deviation in Orchestration, and 22.6% of the overall standard deviation in Artistic Originality.

Since the scoring of the compositions done by the same reviewers may be correlated, we ran a regression analysis that clusters the standard errors by reviewers to yield a more conservative result. Also, the quality of compositions may vary and drive our overall result. We thus ran regression analysis controlling for composition fixed effects to eliminate this concern. All the results are robust using regression analysis with clustered standard errors and additional composition fixed effects controls (See the robustness check section and results in Table 3). The

result is also consistent when we further control for the response time (See the robustness check section).

B. Is there in-group bias based on gender?

To test hypothesis B, we explored whether reviewers have in-group bias toward the same gender. Contrary to the past literature (Nosek et al. 2002; Dovidio and Gaertner 2004; Moss-Racusin 2013) and the majority of the laboratory literature, our findings show no evidence of in-group bias based on gender—reviewers do not give higher scores to composers of their own gender. On the contrary, we found evidence of an out-group bias of male reviewers giving higher ratings to female names. This pattern appears in both general and structured evaluations.

Table 4 shows an opposite sign of in-group bias. Faculty reviewers give statistically significant higher scores to the compositions that are of the opposite gender. This negative coefficient (-0.352) is about 18% of the overall standard deviation. This magnitude is a drastic difference compared with current literature. For example, as stated before, Sandberg (2017) has found that nationalistic in-group bias in the Olympic game is between +7.2% and +23.8% of the overall standard deviation, and our estimated coefficient is in the opposite direction.

This result is mostly driven by the fact that male reviewers favored female composers. Based on Figure 3 and Table 5, male reviewers rated compositions with female names 0.383 points higher in the general rating (0-10 scale) as well as in all five musical dimensions in the structured evaluation, with an average of 0.234 points higher (1-5 scale). This is a sizable opposite-gender bias contrary to the literature—0.383 points correspond to around 20% of the overall standard deviation.

On the other hand, female reviewers were more likely to rate female composers lower in both evaluations, though on a smaller scale. Female faculty rated female composers 0.2 points lower in overall rating (0-10 scale) and an average of 0.09 points lower in the structured evaluation (1-5 scale), though these differences are not statistically significant.

All the above results are presented in Fig. 3. The results are all consistent using regression analysis (See Table 4).

It is worth noting that female faculty gave 0.67 more points than did male faculty in the general evaluation and 0.22 more points in the structured evaluation, regardless of the composer's gender. In other words, female faculty were more generous and relatively unbiased graders, while male faculty tended to give lower scores and favored female composers.

C. Other Characteristics

The data shows that senior faculty (45 years and older) preferred female composers and graded compositions with female names 0.661 points higher than the same compositions with male names in overall evaluation (0-10 scale), equivalent to 34.4% of the overall standard deviation. They also rated female composers higher in all five musical dimensions by an average of 0.276 points (1-5 scale) in the structured evaluation. In contrast, there was no evidence of gender bias

in younger faculty (under 45 years-old) in their evaluations. All the results are consistent using regression analysis and can be found in Fig. 3 (See Table 5).

The participating faculty reviewers differed in tenure status: adjunct instructor, assistant professor, associate professor, and full professor. Bias was only found in two groups: assistant professor and full professor. Both groups preferred female composers in general evaluations and structured evaluations, and compositions associated with female names received higher ratings in all five musical dimensions. All the differences are presented in Fig. 3. The results using regression analysis are nearly identical (See Table 6)

Full professors showed the most significant bias in favor of female composers. Compositions with female names were rated 0.815 points higher on average (0-10 scale) in general evaluation, and an average of 0.393 points higher in structured evaluation (1-5 scale). This bias toward female names in the general evaluation correspond to 42.4% of the overall standard deviation. Assistant professors awarded female composers 0.397 more points in general evaluation ($n = 116$, 0-10 scale; $p = 0.2525$) and an average of 0.259 points higher in structured evaluation (1-5 scale; $p = 0.0925$). Though the overall score using t-test is not statistically significant at 10% significance level, the results using regression analysis are both statistically significant at 10% significance level (See Table 6).

Associate professors tended to rate female composers with a higher score of 0.318 points in general evaluation (under 0-10 scale), but the difference was not statistically significant. Furthermore, no consistent preference for female composers was found in structured evaluation among associate professors. As for evaluations by adjunct faculty, no significant or consistent gender difference in their ratings was found with both evaluation methods. All the results were consistent when further controlled for the female percentage in different ranks.

D. General versus Structured Evaluation

Hypothesis C predicted that discrimination against female composers would be reduced or eliminated by structured evaluation; however, general preference for female composers was found with both evaluation methods. Though the differences in magnitudes are non-comparable due to the measurement designs (0.294 points on 0-10 scale, 0.182 points on 1-5 scale), the results still provide evidence that similar bias persisted in structured evaluation.

In the structured evaluation, gender bias was found in three (tonality/harmony, orchestration, and artistic originality) out of five music categories. We did not find statistically significant differences in tempo/rhythm and form/structure, but the direction is consistent. In-group bias was found in all categories besides tempo/rhythm. That is, reviewers who favored compositions associated with female names rated these pieces higher in almost every aspect of the evaluation: detailed and overall, on specific musical traits and on the complete work. Structured evaluation does not seem to affect faculty reviewers' gender preference.

6. Robustness Check

We conducted further robustness checks using different model specification to see if our results are consistent and valid. Table 8 controls for the response time. We thought it unlikely that response time may matter much to drive the results as some professional judges may simply have downloaded the scores and did the evaluation offline so the response time appears to be short. Similar to our prediction, we find that the results are almost identical to the main result in Table 3. There is also no significant difference in the response time between our randomization group A and group B.

We further controlled for composition fixed effects and respondent fixed effects (See results in Table 8 and Table 9). This specification controls for any difference each composition may have as to compare reviewers' evaluations within the same composition between female names and male names. Controlling for respondent fixed effects can eliminate any concerns of unobservable differences between reviewers as the randomization may not always work perfect. Since our randomization works very well as shown in Table 1, without surprise, the results in Table 8 and Table 9 are consistent with our previous findings.

Finally, one might concern that the results are driven by some outliers. To address this concern, we re-estimate our main coefficients after dropping the observations that have overall scores below 1st and above 99th percentile. The results in Table 11 and Table 12 are consistent with our main results.

7. Discussion and Conclusion

In this study, we replicated a composition competition—an important event in composers' career advancements equivalent to conferences in other fields. The study used new compositions of equal quality by professional composers and employed a simple 0–10 rating scale to evaluate the works. This design help eliminate the potential endogenous unobservables between female and male candidates which is a common issue in observational studies.

This study adds to the discrimination and the in-group bias literature. Contrary to the gender discrimination prediction, our audit study finds a bias favoring female names in professional settings. When we test for the in-group bias hypothesis, we discovered an *out-group* bias, in which male reviewers give higher scores toward female composers. These results were consistent across general and structured evaluations. This gender out-group bias result, while can be surprising to many, is consistent with Bar and Zussman's recently published paper (Bar and Zussman 2020) which shows that, relatively to female testers, male testers are more likely to pass female students' driving tests in Israel. They interpret the result as an extended version of Becker's taste-based model of discrimination (Becker 1961), where instead of incurring distaste toward different race, opposite gender may be preferred in this type of utility-based discrimination model.

Other than out-group bias among male reviewers, we also find faculty members over age 45, assistant professors, and full professors give statistically higher ratings to female composers' works. Among them, full professors showed the most bias favoring female composers. Such heterogeneous effects urge researchers to attend to how institutional norms and policies may, or may not, mediate the gender biases in evaluation. We do not have further evidence to

demonstrate where such heterogeneous effects in gender bias come from. One hypothesis could be a response to the recent and ongoing call for diversity or the #MeToo movement. Clearly, further research in this area is needed to disentangle the patterns and mechanisms of these preferences.

This study speaks to the literature on women's underrepresentation in the academy, which focuses heavily on the STEM fields. Our study fills a critical gap in the literature by examining the extreme case of gender imbalance in a non-STEM field. The puzzling question remains: if we see little gender bias against women, what explains the extreme low representation of women composers? (Q2 Music 2014; O'Bannon 2015, 2016; Brown 2018; Doolittle 2018; Ting 2018). Obviously, our paper alone cannot answer this question, but it can offer some insights for further investigation. First, despite the statistically significant results of our study, the findings should be interpreted with caution. We are reminded by the recent paper on *Journal of Economic Perspective* that discrimination can come in various forms (Small and Pager 2020). The prevalent taste and statistical discrimination models may not capture all reasons behind differential treatment. Discrimination can be unintentional, institutional, historical, or comes through mundane everyday interactions. Our study, focusing on the assessment of sample compositions, holds only one part of composers' career advancements. We did not capture the larger institutional or historical dimension of the problem. In addition, our study does not rule out possible gender bias in other evaluation processes, including the review of resumes, curricula vitae, and letters of recommendation as investigated in other studies (Steinpreis et al. 1999; Bornmann et al. 2007; Knobloch-Westerwick et al. 2013; Moss-Racusin et al. 2012; Trix and Psenka 2003). The other caveat, of course, is that our study was not a hiring experiment, and thus the results have limitations in translating into employment in music composition professions.

Secondly, our findings suggest that women's underrepresentation in academic and professional composition might not be due to unfair evaluation in composition competitions. We found no evidence suggesting that women are discriminated against or receive unfair treatment in competitions, as described by composer Sarah Kirkland Snider (Snider 2017). The results put us in company with other research showing zero or positive bias for females in academia (Williams and Ceci 2015; Ceci and Williams 2014; Breda and Hillion 2016). These researchers argue that the underrepresentation of women in certain STEM fields is less about unfair evaluation in higher education or the labor market, but "pre-college factors and the subsequent likelihood of majoring in these fields." (Ceci et al. 2014) In other words, women's underrepresentation in this field may partially lie on the supply side—in women's decisions not to apply or to pursue such career. Several studies have examined such hypotheses and gendering in music education (Green, 1993, 1994, 1997; Sergeant and Himonides 2016; Legg, 2010; Doolittle 2018), and future research in this area would provide more insights into the explanations of gender imbalance in music composition. In terms of policy implications, our results align with Bagues et al. (2017) that having more female evaluators do not necessarily increase the odds for female candidates to succeed.

Bibliography

Abrevaya, Jason, and Daniel S. Hamermesh. "Charity and favoritism in the field: Are female economists nicer (to each other)?." *Review of Economics and Statistics* 94, no. 1 (2012): 202-207.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. "The impact of jury race in criminal trials." *The Quarterly Journal of Economics* 127, no. 2 (2012): 1017-1055.

Arbuckle, Julianne, and Benne D. Williams. "Students' perceptions of expressiveness: Age and gender effects on teacher evaluations." *Sex Roles* 49, no. 9-10 (2003): 507-516.

Aigner, Dennis J., and Glen G. Cain. "Statistical theories of discrimination in labor markets." *ILR Review* 30, no. 2 (1977): 175-187.

Andrew, Kerry. "Why There Are So Few Female Composers." *The Guardian* (February 3, 2012). https://www.theguardian.com/commentisfree/2012/feb/08/why-so-few-female-composers

Arrow, Kenneth. "The theory of discrimination." *Discrimination in labor markets* 3, no. 10 (1973): 3-33.

Avilova, Tatyana, and Claudia Goldin. "What can UWE do for economics?." In *AEA Papers and Proceedings*, vol. 108, pp. 186-90. 2018.

Bar, Revital, and Asaf Zussman. "Identity and Bias: Insights from Driving Tests." *The Economic Journal* 130, no. 625 (2020): 1-23.

Bagues, Manuel F., and Berta Esteve-Volart. "Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment." *The Review of Economic Studies* 77, no. 4 (2010): 1301-1328.

Bayer, Amanda, and Cecilia Elena Rouse. "Diversity in the economics profession: A new attack on an old problem." *Journal of Economic Perspectives* 30, no. 4 (2016): 221-42.

Becker, Gary S. *The economics of discrimination*. University of Chicago press, 2010.

Bernhard, Helen, Ernst Fehr, and Urs Fischbacher. "Group affiliation and altruistic norm enforcement." *American Economic Review* 96, no. 2 (2006): 217-221.

Bertrand, Marianne, and Esther Duflo. "Field experiments on discrimination." In *Handbook of economic field experiments*, vol. 1, pp. 309-393. North-Holland, 2017.

Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review* 94, no. 4 (2004): 991-1013.

Booth, Alison, and Andrew Leigh. "Do employers discriminate by gender? A field experiment in female-dominated occupations." *Economics Letters* 107, no. 2 (2010): 236-238.

Boring, Anne. "Gender biases in student evaluations of teaching." *Journal of public economics* 145 (2017): 27-41.

Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel. “Gender Differences in Grant Peer Review: A Meta-Analysis.” *Journal of Informetrics* 1 (2007): 226–238. doi:10.1016/j.joi.2007.03.001

Boustan, Leah, and Andrew Langan. "Variation in Women's Success across PhD Programs in Economics." *Journal of Economic Perspectives* 33, no. 1 (2019): 23-42.

Bragger, Jennifer DeNicolis, Eugene Kutcher, John Morgan, and Patricia Firth. “The Effects of the Structured Interview on Reducing Biases Against Pregnant Job Applicants. *Sex Roles* 46, no. 7-8 (April 2002): 215–226. doi:10.1023/A:1019967231059

Broder, Ivy E. "Review of NSF economics proposals: Gender and institutional patterns." *The American Economic Review* 83, no. 4 (1993): 964-970.

Brown, Mark A. “Female Composers Largely Ignored by Concert Line-ups.” *The Guardian* (June 13, 2018). https://www.theguardian.com/music/2018/jun/13/female-composers-largely-ignored-by-concert-line-ups

Buckles, Kasey. "Fixing the Leaky Pipeline: Strategies for Making Economics Work for Women at Every Stage." *Journal of Economic Perspectives* 33, no. 1 (2019): 43-60.

Ceci, Stephen J., and Wendy M. Williams. “Women Have Substantial Advantage in STEM Faculty Hiring, Except When Competing Against More-Accomplished Men.” *Frontiers in Psychology* 6 (2015): 1532. doi:10.3389/fpsyg.2015.01532

Ceci, Stephen J., Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams. "Women in academic science: A changing landscape." *Psychological Science in the Public Interest* 15, no. 3 (2014): 75-141.

Charles, Kerwin Kofi, and Jonathan Guryan. "Prejudice and wages: an empirical assessment of Becker’s The Economics of Discrimination." *Journal of political economy* 116, no. 5 (2008): 773-809.

Chen, Roy, and Yan Chen. "The potential of social identity for equilibrium selection." *American Economic Review* 101, no. 6 (2011): 2562-89.

Chen, Yan, and Sherry Xin Li. "Group identity and social preferences." *American Economic Review* 99, no. 1 (2009): 431-57.

Chen, Yan, Sherry Xin Li, Tracy Xiao Liu, and Margaret Shih. "Which hat to wear? Impact of natural identities on coordination and cooperation." *Games and Economic Behavior* 84 (2014): 58-86.

Currarini, Sergio, and Friederike Mengel. "Identity, homophily and in-group bias." *European Economic Review* 90 (2016): 40-55.

Depew, Briggs, Ozkan Eren, and Naci Mocan. "Judges, juveniles, and in-group bias." *The Journal of Law and Economics* 60, no. 2 (2017): 209-239.

Dickinson, David L., David Masclet, and Emmanuel Peterle. "Discrimination as favoritism: The private benefits and social costs of in-group favoritism in an experimental labor market." *European Economic Review* 104 (2018): 220-236.

Doolittle, Emily. "The Long-Term Effects of Gender Discriminatory Programming." *New Music Box* (June 27, 2018). https://nmbx.newmusicusa.org/the-long-term-effects-of-gender-discriminatory-programming/

Dovidio, John F., and Samuel L. Gaertner. "Aversive racism." *Advances in experimental social psychology* 36 (2004): 4-56.

Fisman, Raymond, Daniel Paravisini, and Vikrant Vig. "Cultural proximity and loan outcomes." *American Economic Review* 107, no. 2 (2017): 457-92.

Folkes, Valerie S., and Shashi Matta. "When Are There Too Many Women? Consumers' Judgments of Gender in Service Groups." *Advances in Consumer Research* 41 (2013): 651–652. http://www.acrwebsite.org/volumes/v41/acr_v41_15521.pdf

Ginsburgh, Victor A., and Jan C. Van Ours. "Expert opinion and compensation: Evidence from a musical competition." *American Economic Review* 93, no. 1 (2003): 289-296.

Goette, Lorenz, David Huffman, Stephan Meier, and Matthias Sutter. "Competition between organizational groups: Its impact on altruistic and antisocial motivations." *Management science* 58, no. 5 (2012): 948-960.

Goldberg, Philip A. "Are Women Prejudiced Against Women?" *Transactions* (April 1968): 28-–30.

Goldin, Claudia, and Cecilia Rouse. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90, no. 4 (September 2000): 715–741. doi:10.1257/aer.90.4.715

Green, Lucy. "Music, Gender and Education: A Report on Some Exploratory Research." *British Journal of Music Education* 10, no. 3 (November 1993): 219–253. doi:10.1017/S0265051700001789

Green, Lucy. "Gender, Musical Meaning, and Education." *Philosophy of Music Education Review* 2, no. 2 (October 1, 1994): 99–105. https://www.jstor.org/stable/40327076?seq=1#page_scan_tab_contents

Green, Lucy. *Music, Gender, Education.* Cambridge, UK: Cambridge University Press, 1997.

Greenwald, Anthony G., and Linda Hamilton Krieger. "Implicit bias: Scientific foundations." *California Law Review* 94, no. 4 (2006): 945-967.

Greenwald, Anthony G., and Mahzarin R. Banaji. "Implicit social cognition: attitudes, self-esteem, and stereotypes." *Psychological review* 102, no. 1 (1995): 4.

Feld, Jan, Nicolás Salamanca, and Daniel S. Hamermesh. "Endophilia or exophobia: beyond discrimination." *The Economic Journal* 126, no. 594 (2016): 1503-1527.

Heckman, James J. "Detecting discrimination." *Journal of economic perspectives* 12, no. 2 (1998): 101-116.

Jannati, Sima, Alok Kumar, Alexandra Niessen-Ruenzi, and Justin Wolfers. "In-group bias in financial markets." (2018). Unpublished manuscript

Kahn, Shulamit, and Donna Ginther. *Women and STEM*. No. w23525. National Bureau of Economic Research, 2017.

Knobloch-Westerwick, Silvia, Carroll J. Glynn, and Michael Huge. "The Matilda Effect in Science Communication: An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest." *Science Communication* 35, no. 5 (February 6, 2013): 603–625. doi:10.1177/1075547012472684

Legg, Robert. "'One Equal Music': An Exploration of Gender Perceptions and the Fair Assessment by Beginning Music Teachers of Musical Compositions." *Music Education Research* 12, no. 2 (2010): 141–149. doi:10.1080/14613801003746592

Leider, Stephen, Markus M. Möbius, Tanya Rosenblat, and Quoc-Anh Do. "Directed altruism and enforced reciprocity in social networks." *The Quarterly Journal of Economics* 124, no. 4 (2009): 1815-1851.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. "Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines." *Science* 347, no. 6219 (January 16, 2015): 262–265. doi:10.1126/science.1261375

Levashina, Julia, Christopher J. Hartwell, Frederick P. Morgeson, and Michael A. Campion. "The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature." *Personnel Psychology* 67, no. 1 (July 10, 2013): 231–293. doi:10.1111/peps.12052

Lundberg, Shelly, and Jenna Stearns. "Women in economics: Stalled progress." *Journal of Economic Perspectives* 33, no. 1 (2019): 3-22.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. "What's in a name: Exposing gender bias in student ratings of teaching." *Innovative Higher Education* 40, no. 4 (2015): 291-303.

McCormick, Lisa. “Higher, Faster, Louder: Representations of the International Music Competition.” *Cultural Sociology*, vol. 3, no. 1 (Mar. 2009): pp. 5–30, doi:10.1177/1749975508100669

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. "Gender bias in teaching evaluations." *Journal of the European Economic Association* 17, no. 2 (2019): 535-566.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. “Science Faculty’s Subtle Gender Biases Favor Male Students.” *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS)* 198, no. 41 (October 9, 2012): 16474–16479. doi:10.1073/pnas.1211286109

Murphy, Michael. “Chen Yi: Composing to Honor Her Past.” *The Choral Journal* 53, no.2 (2012): 28-34.

Nosek, Brian A., Mahzarin R. Banaji, and Anthony G. Greenwald. "Harvesting implicit group attitudes and beliefs from a demonstration web site." *Group Dynamics: Theory, Research, and Practice* 6, no. 1 (2002): 101.

O’Bannon, Ricky. “By the Numbers: Female Composers.” *Baltimore Symphony Orchestra.* https://www.bsomusic.org/stories/by-the-numbers-female-composers/

O’Bannon, Ricky. “What Data Tells us about the 2015-16 Orchestra.” *Baltimore Symphony Orchestra* (December 3, 2015). https://www.bsomusic.org/stories/what-data-tells-us-about-the-2015-16-orchestra-season/

O’Bannon, Ricky. “The Data Behind the 2016-2017 Orchestra Season.” *Baltimore Symphony Orchestra* (2016). https://www.bsomusic.org/stories/the-data-behind-the-2016-2017-orchestra-season/

Paludi, Michele A., and Bauer, William D. “Goldberg Revisited: What’s in an Author’s Name?” *Sex Roles* 9, no. 3 (1983): 387–390. doi:10.1007/BF00289673

Paludi, Michele A., and Strayer, Lisa A. “What's in an Author's Name? Differential Evaluations of Performance as a Function of an Author’s Name.” *Sex Roles* 12, no. 3-4 (February 1985): 353–361. doi:10.1007/BF00287601

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. "Strike three: Discrimination, incentives, and evaluation." *American Economic Review* 101, no. 4 (2011): 1410-35.

Phelps, Edmund S. "The statistical theory of racism and sexism." *The American Economic Review* 62, no. 4 (1972): 659-661.

Price, Joseph, and Justin Wolfers. "Racial discrimination among NBA referees." *The Quarterly Journal of Economics* 125, no. 4 (2010): 1859-1887.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. “How Stereotypes Impair Women’s Careers in Science.” *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS)* 111, no. 12 (March 25, 2014): 4403–4408. doi:10.1073/pnas.1314788111

Sandberg, Anna. "Competing identities: a field study of in-group bias among professional evaluators." *The Economic Journal* 128, no. 613 (2018): 2131-2159.

Shayo, Moses, and Asaf Zussman. "Judicial ingroup bias in the shadow of terrorism." *The Quarterly Journal of Economics* 126, no. 3 (2011): 1447-1484.

Shayo, Moses, and Asaf Zussman. "Conflict and the persistence of ethnic bias." *American Economic Journal: Applied Economics* 9, no. 4 (2017): 137-65.Siegelman, Peter, and J. Heckman. "The Urban Institute audit studies: Their methods and findings." *Clear and Convincing Evidence: Measurement of Discrimination in America, Washington* 187 (1993): 258.

Sergeant, Desmond C., and Evangelos Himonedes. “Gender and the Performance of Music. *Frontiers in Psychology* 5 (2014):276. doi:10.3389/fpsyg.2014.00276

Sergeant, Desmond C., and Evangelos Himonedes. “Gender and Music Composition: A Study of Music, and the Gendering of Meanings.” *Frontiers in Psychology* (March 31, 2016). 10.3389/fpsyg.2016.00411

Snider, Sarah K. “Candy Floss and Merry-Go-Rounds: Female Composers, Gendered Language, and Emotion.” *New Music Box* (May 17, 2017). https://nmbx.newmusicusa.org/candy-floss-and-merry-go-rounds-female-composers-gendered-language-and-emotion/

Small, Mario L. and Devah Pager. “Sociological Perspectives on Racial Discrimination” J*ournal of Economic Perspectives* 34:2 (2020) 49-67

Steinpreis, Rhea A., Katie A. Anders, and Dawn Ritzke. “The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study.” *Sex Roles* 41, no. 7-8 (1999): 509–528. doi:10.1073/pnas.1211286109

Swim, Janet, Barbara Hunter, and Kathryn J. Aikin. “Sexism and Racism: Old-Fashioned and Modern Prejudices.” *Journal of Personality and Social Psychology* (February 1995). doi:10.1037/0022-3514.68.2.199

Tajfel, Henri, Michael G. Billig, Robert P. Bundy, and Claude Flament. "Social categorization and intergroup behaviour." *European journal of social psychology* 1, no. 2 (1971): 149-178.

Ting, Chaowen. “The Data Speaks: Women Composer Representation in the 2016–17 U.S. Collegiate Orchestra Repertoire Selection.” *International Alliance of Women in Music (IAWM) Journal* 24, no. 2 (2018):14–19.

Trix, Frances, and Carolyn Psenka. “Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty.” *Discourse & Society* 14, no. 2 (2003): 191–220. doi:10.1177/0957926503014002277

Watts, Richard. “Career Spotlight: Composer.” *Arts Hub* (May 1, 2018). https://performing.artshub.com.au/news-article/career-advice/performing-arts/richard-watts/career-spotlight-composer-255625

West, Jeremy. "Racial bias in police investigations." (2018). Unpublished manuscript

Whitacre, Eric. “Advice for the emerging composer: Competitions.” (October 25, 2009) http://ericwhitacre.com/blog/advice-for-the-emerging-composer-competitions

Williams, Wendy M., and Stephen J. Ceci. “National Hiring Experiments Reveal 2:1 Faculty Preference for Women on STEM Tenure Track.” *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS)* 112, no. 17 (April 28, 2015): 5360–5365. doi:10.1073/pnas.1418878112

8 Figures and Tables

Figure 1: Graphical analysis of overall evaluation

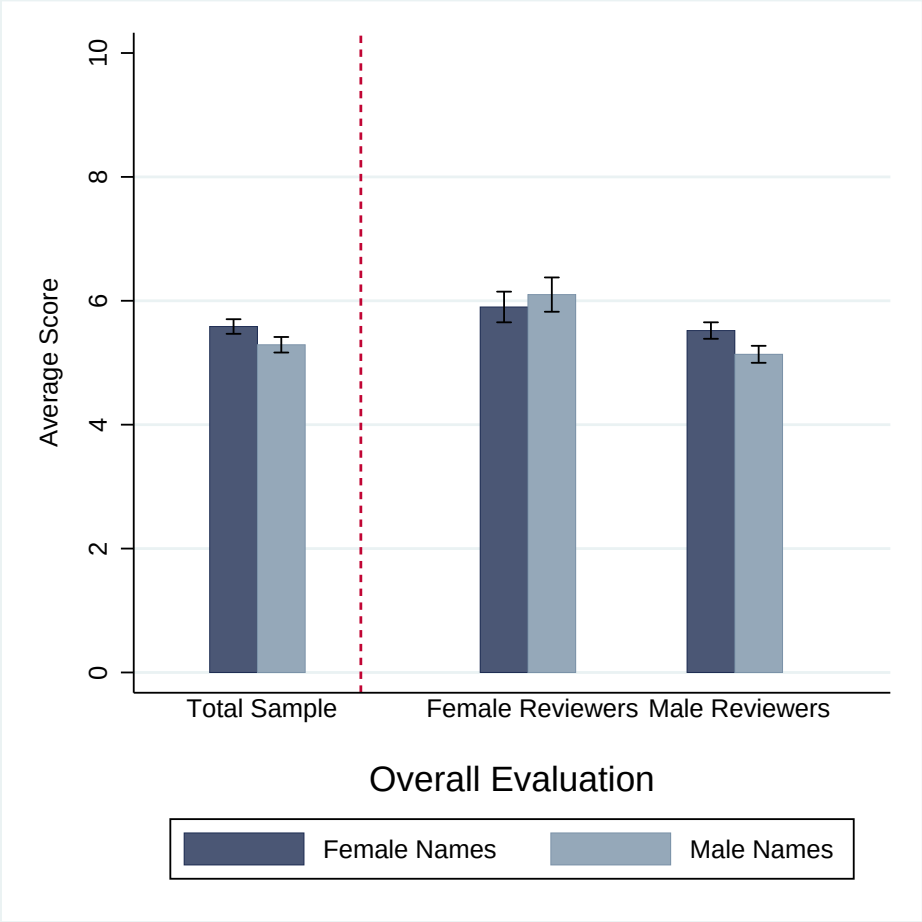


Fig. 1. All the bars represent the average scores in Part I overall evaluation (0-10) for compositions associated with female names vs. male names. The left two bars are mean scores among the total sample; the middle two bars are mean scores among female reviewers; the right two bars are mean scores among male reviewers. Error bars represent standard errors.

Figure 2: Graphical analysis of structured evaluation

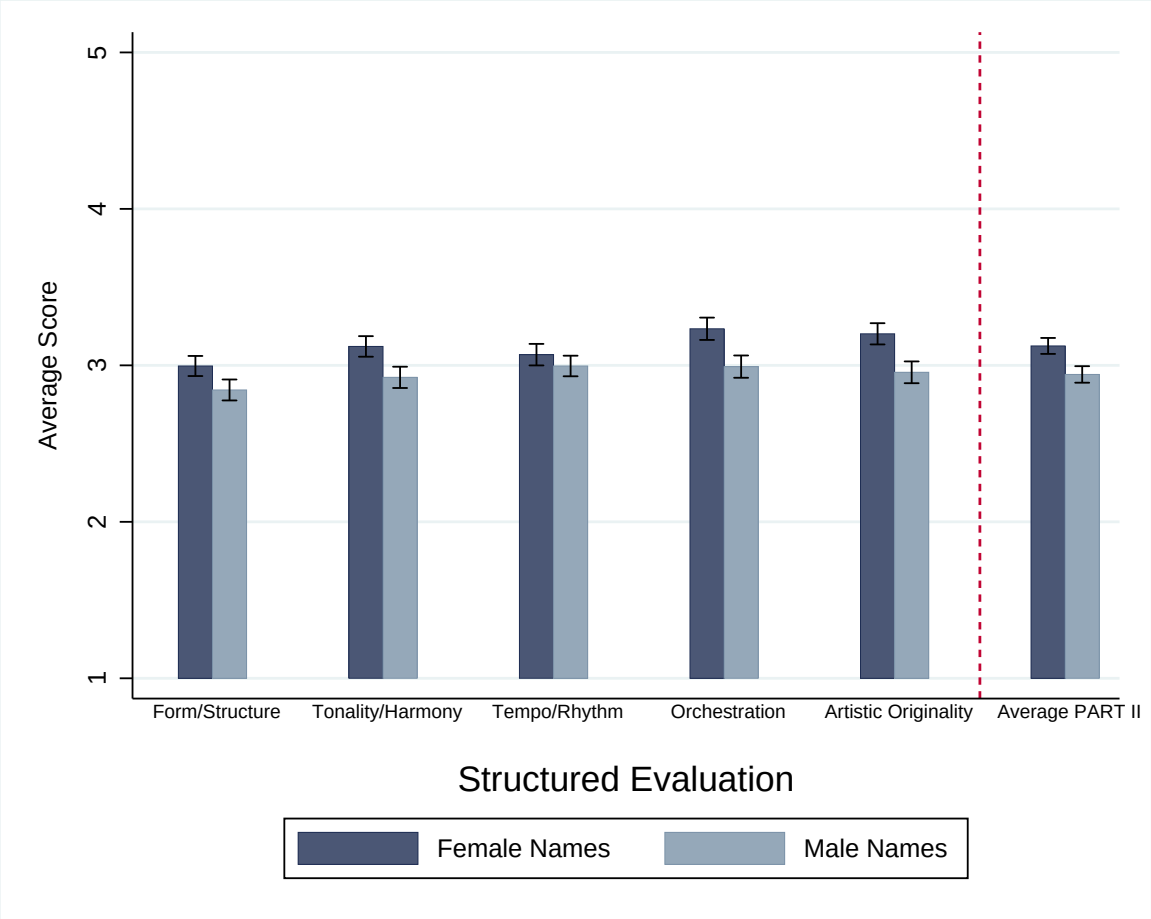


Fig. 2. All the bars represent the average scores in Part II structured evaluation (scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong") for compositions associated with female names vs. male names. Compositions are evaluated in five categories. The last right two bars correspond to the average scores of the five categories. Error bars represent standard errors.

Figure 3: Heterogeneity in gender bias

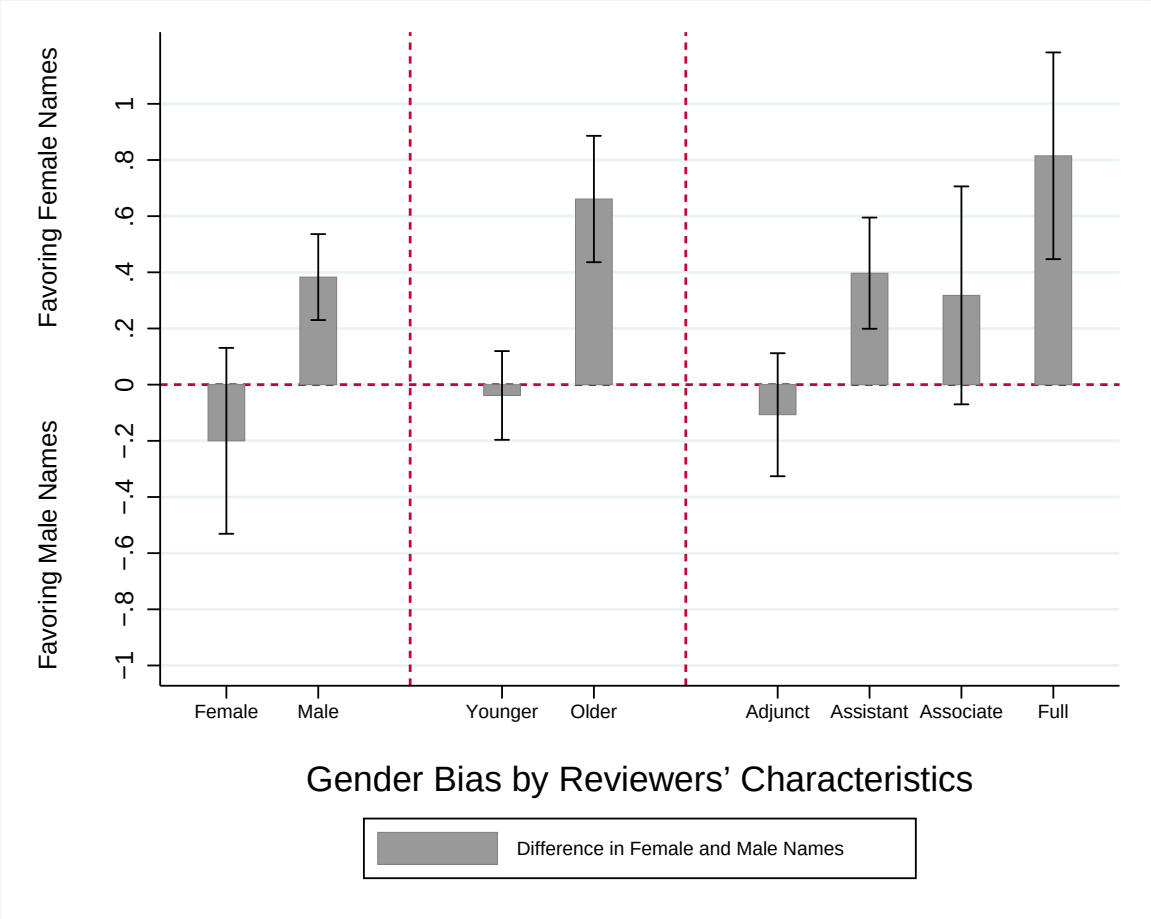


Fig. 3. All the bars represents the gender bias, calcuated as the average score difference between compositions with female names and compositions with male names. Positive (negative) number means higher (lower) score to female names than male names. All the gender bias numbers are estimated based on the regression analysis by different sub-samples based on reviewers' gender, age (45 years old as the cut-off), and rank. (see estimation strategy in the manuscriptand Tables 4-6). Error bars represent standard errors.

Table 1: Summary Statistics and Test for Randomization

	Overall Mean (S.D.) (1)	Randomization A (M,F,M,F) Mean (S.D.) (2)	Randomization B (F,M,F,M) Mean (S.D.) (3)	P-value for testing the difference of (2) and (3) (4)
Female (0/1)	0.16	0.18	0.15	0.60
	(0.37)	(0.39)	(0.36)	
White (0/1)	0.86	0.89	0.84	0.48
	(0.35)	(0.32)	(0.37)	
Age younger than 45 (0/1)	0.52	0.44	0.60	0.07
	(0.50)	(0.50)	(0.49)	
Adjunct professor(0/1)	0.34	0.34	0.33	0.90
	(0.48)	(0.48)	(0.48)	
Assistant professor(0/1)	0.23	0.21	0.25	0.59
	(0.43)	(0.41)	(0.44)	
Associate professor(0/1)	0.18	0.16	0.19	0.70
	(0.38)	(0.37)	(0.40)	
Full professor(0/1)	0.22	0.21	0.22	0.90
	(0.41)	(0.41)	(0.42)	
Public school(0/1)	0.61	0.62	0.60	0.82
	(0.49)	(0.49)	(0.49)	
Average overall evaluation	5.44	5.48	5.39	0.72
	(1.42)	(1.38)	(1.46)	
Number of faculty reviewers	124	61	63	

Note: Standard deviations are in parenthesis. Four faculty reviewers are from the institutions that do not have the conventional rank system, making the four ranks added up <1.

Table 2: Mean Evaluation between Randomly Assigned Female Names and Male Names

	Female names (1)	Male names (2)	Difference (1)-(2) (3)
Part I			
Overall (1-10)	5.585	5.29	0.294*
Part II			
Form/Structure (1-5)	2.996	2.843	0.153*
Tonality/Harmony (1-5)	3.121	2.923	0.197**
Tempo/Rythm (1-5)	3.069	2.996	0.073
Orchestration (1-5)	3.234	2.992	0.242**
Artistic Originality (1-5)	3.202	2.956	0.246**
Average score Part II	3.124	2.942	0.182**
Observation	248	248	496

Note: *** p<0.01, ** p<0.05, * p<0.1. Standard deviations are in parenthesis. Some faculty reviewers are from the institutions that do not have the conventional rank system, making the four ranks not adding up to 1.

Table 3: Overall Results for Gender Bias in Music Evaluation

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female names	0.294**	0.153	0.198**	0.0726	0.242***	0.246***	0.182***
	(0.139)	(0.0935)	(0.0850)	(0.116)	(0.0880)	(0.0857)	(0.0695)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Observations	496	496	496	496	496	496	496
R-squared	0.006	0.005	0.009	0.001	0.012	0.013	0.012

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions.

Table 4: In-group Bias in Music Evaluation

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Same gender	-0.352**	-0.187**	-0.229***	-0.137	-0.224**	-0.229**	-0.201***
	(0.137)	(0.0927)	(0.0845)	(0.116)	(0.0884)	(0.0875)	(0.0695)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Observations	496	496	496	496	496	496	496
R-squared	0.008	0.008	0.012	0.004	0.010	0.011	0.015

Note: All the results are based on regressions with same gender dummy (1 if reviewers and assigned composers are the same gender; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions.

Table 5: Table: Gender Bias in Music Evaluation - Female Heterogeneity

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Female Professor							
Female names	-0.200	-0.125	-0.100	-0.225	0.0250	-0.0250	-0.0900
	(0.331)	(0.196)	(0.218)	(0.211)	(0.205)	(0.229)	(0.154)
Observations	80	80	80	80	80	80	80
R-squared	0.004	0.004	0.003	0.014	0.000	0.000	0.004
Panel B: Male Professor							
Randomly assigned female name	0.383**	0.199*	0.262***	0.136	0.282***	0.291***	0.234***
	-0.153	-0.105	-0.0924	-0.133	-0.0981	-0.0925	-0.0774
Observations	412	412	412	412	412	412	412
R-squared	0.01	0.009	0.015	0.004	0.016	0.018	0.02

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variables. The top panel reports the regression coefficients for female faculty reviewers; The bottom panel reports the regression coefficients for male faculty reviewers. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 123 reviewers' evaluations for four compositions as one reviewer identified the gender question as "other."

Table 6: Table: Gender Bias in Music Evaluation - Age Heterogeneity

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Younger Professor (<45)							
Female names	-0.0385	0.0769	0.138	-0.0692	0.146	0.192*	0.0969
	(0.158)	(0.122)	(0.113)	(0.158)	(0.127)	(0.112)	(0.0803)
Observations	260	260	260	260	260	260	260
R-squared	0.000	0.002	0.004	0.001	0.004	0.008	0.004
Panel B: Older Professor (≥45)							
Randomly assigned female name	0.661***	0.237	0.263**	0.229	0.347***	0.305**	0.276**
	(0.225)	(0.144)	(0.128)	(0.171)	(0.121)	(0.132)	(0.116)
Observations	236	236	236	236	236	236	236
R-squared	0.026	0.013	0.015	0.012	0.025	0.019	0.025

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variables. The top panel reports the regression coefficients for younger faculty reviewers; The bottom panel reports the regression coefficients for older faculty reviewers. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions.

Table 7: Table: Gender Bias in Music Evaluation - Rank Heterogeneity

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Adjunct Professor							
Female names	-0.107	-0.0595	0.179	0	0.155	0	0.0548
	(0.219)	(0.159)	(0.156)	(0.209)	(0.172)	(0.141)	(0.118)
Observations	168	168	168	168	168	168	168
R-squared	0.001	0.001	0.008	0.000	0.005	0.000	0.001
Panel B: Assistant Professor							
Randomly assigned female name	0.397*	0.379**	0.0517	0.224	0.259**	0.379**	0.259**
	(0.198)	(0.183)	(0.128)	(0.187)	(0.110)	(0.157)	(0.0977)
Observations	116	116	116	116	116	116	116
R-squared	0.011	0.032	0.001	0.013	0.013	0.031	0.025
Panel C: Associate Professor							
Randomly assigned female name	0.318	-0.0909	0.159	-0.341	0.409	0.159	0.0591
	(0.388)	(0.189)	(0.228)	(0.244)	(0.242)	(0.208)	(0.154)
Observations	88	88	88	88	88	88	88
R-squared	0.007	0.002	0.006	0.022	0.034	0.005	0.001
Panel D: Full Professor							
Randomly assigned female name	0.815**	0.481*	0.407*	0.315	0.185	0.574**	0.393*
	(0.368)	(0.234)	(0.206)	(0.314)	(0.199)	(0.211)	(0.202)
Observations	108	108	108	108	108	108	108
R-squared	0.043	0.052	0.037	0.019	0.007	0.068	0.048

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variables. The top from the bottom panels reports, respectively, the regression coefficients for adjunct professors, assistant professors, associate professors, and full professors. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 120 reviewers' evaluations for four compositions. Four faculty reviewers are from the institutions that do not have the conventional rank system.

Table 8: Robustness Check: Control for Response Time

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female names	0.294**	0.153	0.198**	0.0726	0.242***	0.246***	0.182***
	(0.139)	(0.0936)	(0.0851)	(0.116)	(0.0880)	(0.0858)	(0.0695)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Extra control for response time	YES	YES	YES	YES	YES	YES	YES
Observations	496	496	496	496	496	496	496
R-squared	0.008	0.006	0.009	0.001	0.012	0.014	0.012

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions.

Table 9: Robustness Check: Alternative Fixed Effects Regression (Gender Bias)

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
First specification							
Female names	0.305**	0.158*	0.203**	0.0860	0.247***	0.246***	0.188***
	(0.127)	(0.0891)	(0.0787)	(0.0889)	(0.0832)	(0.0857)	(0.0608)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Composition fixed effects	YES	YES	YES	YES	YES	YES	YES
Observations	496	496	496	496	496	496	496
R-squared	0.075	0.084	0.088	0.167	0.112	0.059	0.118
Second specification							
Female names	0.305**	0.158	0.203**	0.0860	0.247**	0.246**	0.188***
	(0.147)	(0.103)	(0.0909)	(0.103)	(0.0961)	(0.0990)	(0.0703)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Composition fixed effects	YES	YES	YES	YES	YES	YES	YES
Respondent fixed effects	YES	YES	YES	YES	YES	YES	YES
Observations	496	496	496	496	496	496	496
R-squared	0.618	0.397	0.447	0.464	0.452	0.447	0.495

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions.

Table 10: Robustness Check: Alternative Fixed Effects Regression (In-group Bias)

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
First specification							
Same gender	-0.352**	-0.187**	-0.229***	-0.137	-0.224**	-0.229**	-0.201***
	(0.137)	(0.0927)	(0.0845)	(0.116)	(0.0884)	(0.0875)	(0.0695)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Composition fixed effects	YES	YES	YES	YES	YES	YES	YES
Observations	496	496	496	496	496	496	496
R-squared	0.008	0.008	0.012	0.004	0.010	0.011	0.015
Second specification							
Same gender	-0.380***	-0.200*	-0.251***	-0.184*	-0.245**	-0.250**	-0.226***
	(0.145)	(0.103)	(0.0903)	(0.102)	(0.0965)	(0.0996)	(0.0693)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Composition fixed effects	YES	YES	YES	YES	YES	YES	YES
Respondent fixed effects	YES	YES	YES	YES	YES	YES	YES
Observations	496	496	496	496	496	496	496
R-squared	0.621	0.400	0.452	0.470	0.451	0.447	0.500

Note: All the results are based on regressions with same gender dummy (1 if reviewers and assigned composers are the same gender; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions.

Table 11: Robustness Check: Dropping Outliers (Gender Bias)

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female names	0.333**	0.169	0.210**	0.102	0.250**	0.255**	0.197***
	(0.150)	(0.104)	(0.0904)	(0.103)	(0.0967)	(0.101)	(0.0711)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Composition fixed effects	YES	YES	YES	YES	YES	YES	YES
Respondent fixed effects	YES	YES	YES	YES	YES	YES	YES
Observations	492	492	492	492	492	492	492
R-squared	0.611	0.395	0.448	0.465	0.451	0.447	0.495

Note: All the results are based on regressions with female assigned name dummy (1 if assigned with female names; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions. In this robustness check, we dropped observations that have overall scores below the 1st and above the 99th percentile.

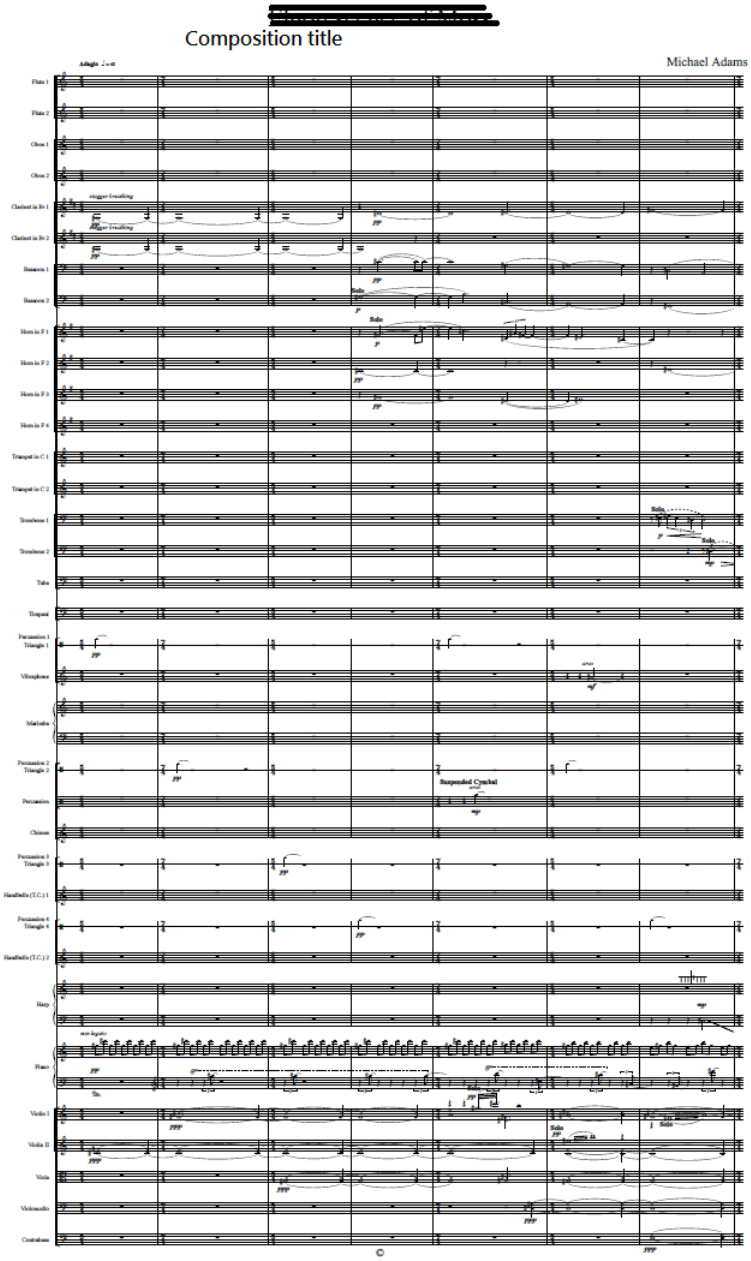
Table 12: Robustness Check: Dropping Outliers (In-group Bias)

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Same gender	-0.385**	-0.194*	-0.241***	-0.186*	-0.235**	-0.253**	-0.222***
	(0.148)	(0.104)	(0.0902)	(0.103)	(0.0973)	(0.101)	(0.0705)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Composition and respondent fixed effects	YES	YES	YES	YES	YES	YES	YES
Respondent fixed effects	YES	YES	YES	YES	YES	YES	YES
Observations	492	492	492	492	492	492	492
R-squared	0.614	0.397	0.451	0.470	0.449	0.447	0.498

Note: All the results are based on regressions with same gender dummy (1 if reviewers and assigned composers are the same gender; 0 otherwise) as the independent variable and all the scores in Part I and Part II as the outcome variable. The top row represent the regression coefficient. Robust standard errors clustered with reviewers are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The number of observations is the unique combination of composition and reviewer, based on 124 reviewers' evaluations for four compositions. In this robustness check, we dropped observations that have overall scores below the 1st and above the 99th percentile.

9 Appendix A: Figures and Tables

Figure A-1: One Example Page from The Score



10 Appendix B: Composition Evaluation Survey

Start of Block: Welcome Message

PART I Thank you for participating in our survey. We appreciate your time and expertise.

The purpose of this study is to better understand criterion used in assessing musical compositions. Below you will be asked to evaluate four (4) 10-min orchestral works. Please evaluate the compositions as if you were judging a final round of a Call for Score composition competition.

The survey takes approximately 40-60 minutes to complete. You can save your answers and return at a later time. Upon completion of the survey, you will be awarded a $75 gift card for your time.

Click the next button to get started!

End of Block: Welcome Message

Start of Block: (B1) Music Evaluation: General

Q1 Please evaluate the composition as if you were judging a composition competition, and provide your general recommendation for composition No.1.

Fluorescence of Moss_Score.pdf
Fluorescence of Moss_Recording.mp3

	0	1	2	3	4	5	6	7	8	9	10
Overall Recommendation ()											

Page Break

Q2 Please evaluate the composition as if you were judging a composition competition, and provide your general recommendation for composition No.2.

Stockholm_Score.pdf
Stockholm_Recording.mp3

	0 1 2 3 4 5 6 7 8 9 10
Overall Recommendation ()	

Page Break

Q3 Please evaluate the composition as if you were judging a composition competition, and provide your general recommendation for composition No.3.

The Irresistible Embrace of Singularity_Score.pdf
The Irresistible Embrace of Singularity_Accompanying Tape.mp3
The Irresistible Embrace of Singularity_Recording.mp3

	0 1 2 3 4 5 6 7 8 9 10
Overall Recommendation ()	

Page Break

Q4 Please evaluate the composition as if you were judging a composition competition, and provide your general recommendation for composition No.4.

FiddleSticks!_Score.pdf
FiddleSticks!_Recording.mp3

	0 1 2 3 4 5 6 7 8 9 10
Overall Recommendation ()	

End of Block: (B1) Music Evaluation: General

Start of Block: (B2) Music Evaluation: Specifics

PART II In the next section, you will be asked to provide detailed assessments of the four pieces as we are interested in understanding the criterion used in evaluating musical compositions. Please refer to the scores and recordings when providing your evaluation.

Page Break

Q5 For Musical Composition No.1, please rate the artistic achievement of the following:

Fluorescence of Moss_Score.pdf
Fluorescence of Moss_Recording.mp3

	Extremely Weak (1)	Somewhat Weak (2)	Neutral (3)	Somewhat Strong (4)	Extremely Strong (5)
Form / Structure (1)	○	○	○	○	○
Tonality / Harmony (2)	○	○	○	○	○
Tempo / Rhythm (3)	○	○	○	○	○
Orchestration (4)	○	○	○	○	○
Artistic Originality (5)	○	○	○	○	○

Page Break

Q6 For Musical Composition No.2, please rate the artistic achievement of the following:

Stockholm_Score.pdf
Stockholm_Recording.mp3

	Extremely Weak (1)	Somewhat Weak (2)	Neutral (3)	Somewhat Strong (4)	Extremely Strong (5)
Form / Structure (1)	○	○	○	○	○
Tonality / Harmony (2)	○	○	○	○	○
Tempo / Rhythm (3)	○	○	○	○	○
Orchestration (4)	○	○	○	○	○
Artistic Originality (5)	○	○	○	○	○

Page Break

Q7 For Musical Composition No.3, please rate the artistic achievement of the following:

The Irresistible Embrace of Singularity_Score.pdf
The Irresistible Embrace of Singularity_Accompanying Tape.mp3
The Irresistible Embrace os Singularity_Recording.mp3

	Extremely Weak (1)	Somewhat Weak (2)	Neutral (3)	Somewhat Strong (4)	Extremely Strong (5)
Form / Structure (1)	○	○	○	○	○
Tonality / Harmony (2)	○	○	○	○	○
Tempo / Rhythm (3)	○	○	○	○	○
Orchestration (4)	○	○	○	○	○
Artistic Originality (5)	○	○	○	○	○

Page Break

Q8 For Musical Composition No.4, please rate the artistic achievement of the following:

FiddleSticks!_Score.pdf
FiddleSticks!_Recording.mp3

	Extremely Weak (1)	Somewhat Weak (2)	Neutral (3)	Somewhat Strong (4)	Extremely Strong (5)
Form / Structure (1)	○	○	○	○	○
Tonality / Harmony (2)	○	○	○	○	○
Tempo / Rhythm (3)	○	○	○	○	○
Orchestration (4)	○	○	○	○	○
Artistic Originality (5)	○	○	○	○	○

Page Break

Q9 Finally, please make your final recommendation for the winner of the mock competition. (pick only one piece)

- ○ Fluorescence of Moss (Michael Adams) (1)
- ○ Stockholm (Rebecca Moore) (2)
- ○ The Irresistible Embrace of Singularity (Sean Campbell) (3)
- ○ FiddleSticks! (Tara Davis) (5)

End of Block: (B2) Music Evaluation: Specifics

Start of Block: Demographics

Q10 Please specify your age range.

- ○ 18-24 years old (1)
- ○ 25-34 years old (2)
- ○ 35-44 years old (3)
- ○ 45-54 years old (4)
- ○ 55-64 years old (5)
- ○ 65-74 years old (6)
- ○ 75 years or older (7)

Q11 Please specify your gender.

- ○ Female (1)
- ○ Male (2)
- ○ Other (3)

Q12 Please specify your ethnicity.

- ○ White (1)
- ○ Black or African American (2)
- ○ American Indian or Alaska Native (3)
- ○ Hispanic or Latino (4)
- ○ Asian/Pacific Islander (5)
- ○ Other (6)

End of Block: Demographics