Gender Bias in Competitive Music Composition Evaluation: An Experimental Study

By CHAOWEN TING, YATING CHUANG, JOHN CHUNG-EN LIU *

February 10, 2022

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. We conducted an audit experiment to examine whether such gender imbalance is due to unfair judgment in evaluative settings. We invited composition faculty to rate new compositions with randomly assigned gendered names along with live recordings directed by the same conductor. Contrary to our hypotheses, we do not observe gender bias against women. Instead, there is evidence that compositions associated with female names are rated higher than those with male names. There is no evidence of in-group bias either: reviewers do not favor compositions from composers of their genders. In the heterogeneity analysis, we find suggestive evidence that male faculty and senior faculty favor female composers in both general and structured evaluations.

* Chuang: Department of Economics, National Taipei University, New Taipei City, 23741 (email: yating@gm.ntpu.edu.tw); Ting: School of Music, Georgia Institute of Technology, Atlanta, GA 30332 (email: chaowen.ting@music.gatech.edu); Liu: Department of Sociology, National Taiwan University, Taipei, 10617 (email: chungenliu@ntu.edu.tw).

We thank College of design, Georgia Institute of technology, and Occidental College for generous funding support. For helpful comments, we thank Jason Freeman, Tiffany Yeh, Wilbur Lin, Yuan-Chen Li, Elaine Liu, and seminar participants at National Taiwan University, National Taipei University, National Chengchi University, Fu Jen Catholic University, National Central University, National Central University, and National Tsing Hua University.

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 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 led to lively discussion about gender atmosphere in the economics profession (Wu 2018; Wu 2020) and stimulated new research on gender bias in publication in economics journals and conferences (Hengel, 2018; Hospido and Sanz, 2019; Chari and Goldsmith-Pinkham, 2017; Card et al., 2020). 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 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, 2018). Either way, women face disadvantages and are less likely to advance in their careers.

[1] The National Science Foundation and National Center for Education Statistics. https://nces.ed.gov/programs/coe/indicator_csc.asp
[2] Bayer and Rouse (2016) provides a review in this demand side explanation for understanding diversity in the economics profession.

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 (O'Bannon, 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).

Besides the numbers, previous research has demonstrated the existence of systematic gender discrimination in music composition profession, to the extent that female composers need to manage their identities to mitigate such disadvantage (Bennett et al. 2018, Bennett et al. 2019, Cannizzo and Strong 2020). Gender bias in music evaluations (Goldin & Rouse, 2000; Leonard, 2007; Miller 2016), as well-documented in the literature, is certainly a component of gender discrimination. Against this background, we have witnessed more conversations about gender and diversity issue in composition and some small steps were made (Robin, 2017; McClary and de Boise 2019). For example, more female composers rose to prestigious programs in recent years. The year 2017 was the historical year when all three finalists of the Pulitzer Prize for Music were women. We also see social movement sprouting from social media, such as the hashtag #HearAllComposers, to promote diversity in music. 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 the evaluations differ in more structured evaluation methods, such as giving detailed ratings on various musical dimensions. 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 (NASM) 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.

Our paper contributes 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 Bertrand and Duflo, 2016). The leading economic models to explain discrimination are (1) Becker's taste-based model (Becker, 1957; Charles and Guryan, 2008), (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 perception of skill distribution between different groups, and predicts employers or evaluators would favor the majority groups (e.g., white and male) because of their perceived higher productivity. Recent literature on the belief-based models of discrimination is also relevant (Bohren et al. 2019). In this model, there is discrimination toward the minority group in the initial stage, but when reviewers receive new information and the minority group gains more reputation, evaluators would update their beliefs and thus discrimination would disappear or even reverse. In these models, we often observe 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 difficult to examine the mechanism to explain the sources of discrimination. In our contexts, we examine gender bias in a unique field with large gender imbalance, and are able to observe the identities of evaluators and evaluatees.

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, 2018). 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 the 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 (Fisman 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 (Feld et al., 2016), 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 are 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 and Mullainathan (2004) (See more review on field experiments in testing discrimination by Bertrand and Duflo (2017)). In Goldin and Rouse's (2000) examination of orchestra audition outcomes—the context mostly similar to ours, 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 ensure we select the appropriate samples. We have both 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, 2018). 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, 1993; 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 are 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 the recruiter's race or gender. Even in audit studies, the trained interviewers 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 confounding factors 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 that compositions associated with female names are rated higher than compositions associated with male names. This pattern also persists in the more structured evaluations when participants assess compositions in different subcategories. We also find some evidence of an *out-group* bias by male evaluators in favor of female composers in the music composition. In addition, senior faculty shows preference toward female composers than more junior faculty. Our results align with various studies that fail to find male 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, the common in-group bias hypothesis may not hold outside laboratories as this preference may differ in contexts. Second, even in a highly professional setting where gender imbalance persists, it is possible for evaluators to update their beliefs and reverse the gender bias. Finally, biases in treating the minority group may not be an essential explanation for the underrepresentation of women in the music composition profession. This finding urges us to also look closely at the "supply side" factors, including self-confidence, stereotypes, training pipelines, etc., to understand the underlying causes of inequalities (Bordalo et al, 2016; Bordalo et al, 2019).

[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. Sections 3 and 4 introduce the experimental design and the estimation strategy. Section 5 discusses the results. Section 6 presents the robustness checks. Section 7 discusses the findings. Section 8 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 a significant impact on the later success. Though in a different context, Ginsburgh and Van Ours (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 sends out invitations to invite professional composers as reviewers. The scores, often along with a recording, are uploaded to a platform and the reviewers will give evaluations of the compositions. 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. In theory, we could create a control group without any composer names as the baseline; in practice, such anonymous design deviates from common practices in the profession. Various esteemed and highly influential composition competitions, such as Guggenheim, Fromm, Koussevitzky, ASCAP, Barlow, are not anonymous, which we follow suit. Moreover, as our study focuses on a group of highly specialized professionals, it is uniquely challenging to find qualified subjects. We seek to maximize the statistical power by not adding another control group. Since our main purpose is to understand the relative level of gender bias instead of the absolute level of bias, we keep this standard practice with composers' names included.

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., 2013). 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 the 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. 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.

With limited gender cues, respondents in our pre-tested qualitative survey interviews were able to correctly recall the gender of the composers, indicating that subjects were truly "treated" in this design. Also, because this population is a highly white-dominated discipline, none of the respondents have conveyed any racial concerns toward the names. We also test those names using NamePrism (Ye et al., 2017), a non-commercial ethnicity/nationality classification tool and have been used in many economics papers (Diamond et al., 2019; Tang, 2020; Honigsberg and Jacob, 2021; Kempf and Tsoutsoura, 2021). Our pseudo names are unambiguously (all are around 90%) classified as white names.

<u>Recruitment</u>

Faculty members were recruited from 377 music institutions that participated in the 2013-14 National Association of Schools of Music (NASM) 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 B: Survey Materials for reference) via Qualtric system through a personalized survey link in the invitation email. This study has been approved by the Institutional Review Board (IRB) at Georgia Institute of Technology and the subjects were informed in the email invitation (Appendix B). The subjects would give consent when they filled out the online survey. The reviewers 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. We use personalized links to ensure data quality and prevent 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 sample may still differ in other unobservable dimensions. Nonetheless, the professionals in the music field here can still provide meaningful insights as this setting has never been examined in the discrimination literature. 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 group 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 group B. As Table 1 shows, our randomization works well since none of the differences between the 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. Some economists have concerns over "deception" in audit studies and have proposed using new experimental methods to elicit preference without deception (Kessler et al., 2019). We argue that it is less of a problem in our case. The first-order problem in resume studies is whether the fake resumes differ from the real resumes. We do not have this issue since all of our composition pieces are not “fake”—they are written by actual composers. Another concern is on the “callback rate” in the resume studies, in which researchers worry that the employers’ evaluation of the resume may confound with their expectation of the candidates’ acceptance. For example, callback rates are usually higher among unemployed candidates than employed candidates, not necessarily reflecting the employer’s preference. In our study, we are not subject to this problem as we ask evaluators to give scores rather than make phone calls.

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 detail. Presumably, this structured evaluation would prompt reviewers to consider 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– 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 B.

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 the reviewer. Our estimation equation is:

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

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. δ_c controls for composition fixed effects; γ_i controls for reviewer fixed effects. This model 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 reviewer fixed effects can eliminate any concerns of unobservable differences between reviewers.

[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.

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. Since the scoring of the compositions done by the same reviewers may be correlated, we cluster the standard errors by reviewers.

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:

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

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, such as composition fixed effects (δ_c) and reviewer fixed effects (γ_i). Similar to the previous specification, we cluster the standard errors by reviewers.

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 the general rating and structured rating (See Figure 1). Examining the overall ratings by gendered names, we find that female composers scored on average 0.305 points higher on a 0-10 scale (See Figure 1 and Table 2). This differential bias is non-negligible—about 16% 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 2018). 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 0.188 more points on a 1-5 scale (See Figure 2 and Table 2). Specifically, the difference was 0.158 in Form / Structure, 0.203 in Tonality / Harmony, 0.086 in Tempo / Rhythm, 0.247 in Orchestration, and 0.246 in Artistic Originality (statistically significant at 0.05 significance level except for Form and Tempo and; See Figure2 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.

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, 2012) 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 3 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 (-0.38) is almost 20% of the overall standard deviation. This magnitude is a drastic difference compared with current literature. For example, as stated before, Sandberg (2018) 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 with similar magnitude.

This result is mostly driven by the fact that male reviewers favored female composers. Based on Table 4, male reviewers rated compositions with female names 0.402 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 0.244 points higher (1–5 scale). This is a sizable opposite-gender bias contrary to the literature—0.402 points correspond to around 21% 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 0.278 points lower than male composers in overall rating (0–10 scale) and an average of 0.141 points lower in the structured evaluation (1–5 scale).

We further test the rating difference between female and male faculty (See the interaction effect in Panel C Table 4). The difference in overall rating between female and male faculty reviewers is marginally statistically significant (p-value<0.1); yet this difference is sizable. The difference in the average scores in the structured rating (column 7) between female and male faculty reviewers is statistically significant (p-value<0.01).

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 (both are statistically significant), 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. Effect across the Distribution of Overall Evaluation

We follow Kessler et. al (2019) to implement a counterfactual callback exercise. At each overall evaluation scoring level, we can generate an equivalent counterfactual callback rate treating each scoring level as the callback threshold (composers would be "called back" if their overall evaluation score is above that given score as if in a hiring situation). Even though our study is not a hiring experiment, our granular measure—overall evaluation score—is advantageous to explore whether evaluators' preferences would differ in the tails of the distribution. Unlike typical audit studies, the binary callback rate variable can be sensitive to the low callback rate environment (Kessler et al. 2019).

Two analytical tools are used here. The first graph is the empirical cumulative distribution function (CDF) of the overall evaluation scores. In our context, we will compare the CDFs of female composers and male composers. The second exercise is to generate counterfactual callback rates by assuming that evaluators would call back the composers if their overall evaluation scores were at or above a threshold. We can then observe the differences in callback rates across each scoring threshold through a linear probability model.

We plot the CDF of the ratings of hypothetical male and female composers (see Figure 3 panel a); We also show the differences in the counterfactual callback rates (Figure 3 panel b). We see that, in Figure 3(b), the coefficients are higher than zero at almost all levels of selectivity (less pronounced at the tail), meaning that our result is almost all qualitatively consistent across the distribution of the overall score. We further employ the same analysis by reviewers' gender in Figure 4. When reviewers are female, the difference in the counterfactual callback rates between female composers and male composers are nearly zero at all levels of the scoring threshold (Figure 4 Panel b) On the contrary, when reviewers are male, callback rates for female composers are consistently higher than their male counterparts across the whole scoring distribution (Figure 4 Panel d), although this difference is less pronounced at the tail. This additional analysis reassures that our study is well powered and generalizable—compositions associated with female names have an impact on reviewers' preference almost throughout the distribution of the overall evaluation scores.

D. Other Characteristics

The data shows that senior faculty (45 years and older) preferred female composers and graded compositions with female names 0.536 points higher than the same compositions with male names in overall evaluation (0–10 scale), equivalent to 28% of the overall standard deviation. They also rated female composers higher in all five musical dimensions by an average of 0.205 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 Table 5 (See also Appendix Figure A-1). We further test whether there is statistically significant difference in evaluation between younger faculty and elder faculty, Panel C in Table 5 shows that this difference is only at the border line of statistically significant in overall evaluation (P-value<0.1), but not in other structured evaluation.

Table 6 presents results among participating faculty reviewers differed in tenure status: adjunct instructor, assistant professor, associate professor, and full professor. Bias was found among full professors. They preferred female composers in general evaluations and structured evaluations, and compositions associated with female names received higher ratings in all five musical dimensions (although dimensions tempo and orchestration are not statistically significant.). All the differences are presented in Table 6 (See also Appendix Figure A-1). Full professors showed the most significant bias in favor of female composers. Compositions with female names were rated 0.852 points higher on average (0–10 scale) in general evaluation, and an average of 0.415 points higher in structured evaluation (1–5 scale). This bias toward female names in the general evaluation corresponds to 44% of the overall standard deviation. All the lower-ranked professors, including adjunct professors, assistant professors, and associate professors do not have statistically significant bias toward women. We further test whether there is a statistically significant difference in evaluation between tenured professors (Associate professor and above) and others, Panel C in Table 6 shows that this difference is only at the border line of statistically significant in overall evaluation (P-value<0.1), but not in other structured evaluation. It is also worth noting that the patterns between the overall score and the average score (based on the structured evaluation) seem slightly different in Tables 5 and 6. We suspect that evaluators may have different opinions concerning what considers to be a good composition. We explain more details about the general evaluation using the overall score versus the structured evaluation in the next section.

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 the overall evaluation and all five music categories (although form/structure and tempo/rhythm are significant at 10% level). That is, reviewers who favored compositions associated with female names rated these pieces higher in 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. One may notice a small discrepancy between overall scores and the average scores of structured evaluation in the heterogeneity analysis. We attribute such discrepancy to the low statistical power, making the results less stable.

One caveat of our study is that we use a within-subjects test instead of a between-subjects test when comparing general versus structured evaluation. In an ideal situation, we can assign some respondents to conduct only the overall evaluation survey, and others to conduct only the structured ones to create a between-subjects test. Nevertheless, we designed a within-subjects test for several reasons. First, with this unique population, we will be even further constrained by the statistical power if using a between-subjects test. Second, we could not ask respondents to evaluate only a structured evaluation as such evaluation is not conventional in the composition competition. Third, we could not ask reviewers to evaluate only the overall evaluation either because this setting may at odds with our stated invitation purpose to study the criteria of evaluation.

We recognize our design limitation that the reviewers may remember their overall evaluation scores and rate the structured evaluation to be consistent with their initial assessment. We make such harmonization more difficult by making it impossible to "go back to the previous page" to change their overall scores once the reviewers are in the structured section. In addition, the scales of the scores are different (i.e. overall evaluation (1-10) and structured evaluation (1-5)), making it more challenging to intentionally "fit" the two types of assessment.

6. Robustness Check

We conducted further robustness checks using different model specifications to see if our results are consistent and valid. Table 7 deals with 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. To test whether the results are robust, we simply drop those who have unusual short response time (less than 10% of the bottom tail). Similar to our prediction, we find that the results are almost identical to the main result in Table 2. There is also no significant difference in the response time between our randomization group A and group B.

Furthermore, to assure our external validity, we coded the background information of more than one thousand respondents and non-respondents, and test if they are different from each other. We are constrained that only basic information is available from the faculty's website. Based on the results in Table 8, we find that there our respondents are very similar to non-respondents in terms of the proportion of gender, and whether the faculty is from a public or private school. We find that faculty members in our experiment are less likely to be tenured faculty. We do not know if this response rate difference is correlated with gender bias. Nonetheless, if our heterogeneity analysis holds (i.e. senior faculty favors female composers), our current results may underestimate the female favoritism.

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 9 and Table 10 are consistent with our main results.

7. Interpretation

To make sense of our findings, we seek to interpret the mechanism behind our observed results with different economic models of discrimination. In the previous literature, most papers distinguish between taste-based discrimination model and statistical discrimination model. The newer strand of literature using biased belief to explain discrimination is also relevant in our contexts (Fryer, 2007; Fryer and Jackson, 2008; Schwartzstein 2014; Bohren et al., 2019). Our results are consistent with an extended version of taste-based discrimination model (i.e. the in/out-group bias). While our results are less in line with the predictions from the classical statistical discrimination model, the findings can relate to the "belief-based" model, in which a belief update at a later stage could correct or even reverse the initial bias. We, however, have limited information to adjudicate the two mechanisms. Let us examine the two mechanisms in detail below.

7.1 Taste-based Discrimination and in/out-group bias

The taste-based discrimination mechanism started from Becker's utility-based discrimination model (Becker, 1957; Charles and Guryan, 2008), which illustrates that employers may prefer a certain type of group (e.g. white over black employees) given the same ability because employers incur disutility from interacting with that certain group. Empirically, in-group or out-group bias can be seen as an extended version of Becker's model. For example, Bar and Zussman (2020) found that, in the driving tests, male testers generate utilities from interacting with students of the opposite gender, resulting in a higher passing rate among female students. In this regard, our results may be due to a similar taste-based gender bias, which male reviewers incur utilities by potentially having more female composers winning awards, and thus prefer female composers in the experiment.

7.2 Statistical Discrimination and the belief-based model

The second common model to explain discrimination is statistical discrimination. From this perspective, our results would imply that reviewers view female composers, based on previous information, as more capable composers, and thus rate them higher. Such explanation is unlikely as we have ample evidence showing the opposite. We therefore argue that our findings are less in line with the predictions from the classical statistical discrimination model.

Our experiment, however, does not rule out mechanisms related to statistical discrimination against women. In the more recent "belief-based" literature, discrimination is dynamically updated with new information. For example, there is evidence of settings where evaluators hold initial discriminatory beliefs against a group but quickly update them if they see evidence of good performance. Discrimination could be reduced or even reversed after the discriminated group earns more reputation. Linking this literature to our results, it is possible that the evaluators know how difficult it is to "make it" in the music world as a female composer. That is, a previous stage of selection, as well as a “belief update,” has already occurred. In the evaluators’ updated belief, these compositions are worthy of higher ratings as are authored by women who choose to become music composers. In a similar vein, our suggestive evidence that senior faculty are more favorable of female composers could also be related to this “belief update” mechanism. [5]

The caveat, however, is that all the compositions we selected are commissioned pieces with our collaborated institution, these composers may have already made it to the “later stage.” We do not run evaluations among more junior level composers and thus cannot directly observe the belief update dynamics.

Finally, we also want to mention that our study is relevant to the literature on joint versus separate evaluation. Bohnet et al. (2016) find that—with an “evaluation nudge”—jointly evaluated resumes show less implicit bias against women than sequential evaluations. Given that our experiment treats every reviewer with two male names and two female names, the design could potentially nudge the reviewers to mitigate their biases. We could not explore this mechanism thoroughly due to our experiment design, but want to draw attention to this feature for future studies.

8. Conclusion

[5] Some may be curious whether there are “gendered features” in the music compositions to drive our results. Based on the music experts’ opinion, there is no consensus that compositions can sound more “male” or “female.” Second, for the composers’ actual identities, there are two pieces written by males and two pieces written by females. Our current result includes composition fixed effects in the model, so this factor has been controlled for.

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 helps 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.

Some may worry that our results are driven by the experiment demand effect. For example, faculty members may google the name of the composition and figure out that this is a study about gender, resulting in Hawthorne effect with male evaluators giving favored evaluations toward female composers. We argue that this effect is unlikely. Those compositions are all newly commissioned works with no public info available online, so the actual composers' names cannot be revealed. We also double-checked that our fictitious names do not coincide with real composers. It is quite common that junior composers or composition students do not yet have a prominent online portfolio. In addition, one may worry about the two-female-two-male design in our experiment, leading to a subtle signaling effect for diversity. We cannot fully eliminate this factor. Nonetheless, this design is not uncommon as many female composers advance to the finalists of multiple famous competitions in recent years.

All in all, 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 extremely low representation of women composers? (O'Bannon, 2014, 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. As for a concrete policy implication, our results align with Bagues et al. (2017) that having more female evaluators does not necessarily increase the odds for female candidates to succeed.

Despite the statistically significant results of our study, the findings should be interpreted with caution. We are reminded by the recent paper in the *Journal of Economic Perspectives* 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. Finally, our study cannot capture temporal changes in this field. If the recent call for diversity is an explanation of our results, it might only “benefit” a small number of composers of a generation, thus the larger inequality remains.

Keeping in mind the limitations, our findings suggest that women’s underrepresentation in professional composition might not be due to unfair evaluation in composition competitions, or at the very least, the demand side discrimination should not be seen as the only cause. Women’s underrepresentation in this field may partially lie on the supply side—in women’s decisions not to apply or to pursue such careers. Our results can speak to other research showing zero or positive bias for females in academia (Ceci and Williams, 2015; Williams and Ceci, 2015; Breda and Hillion, 2016), in which 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). To what extent our insights can be applied to STEM fields and other professions remains an open question. Much further research is required in other aspects of this profession, such as composers’ employment, pay, and career advancement, to arrive at a more conclusive judgment.

References

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.

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

Antonovics, Kate, and Brian G. Knight. "A new look at racial profiling: Evidence from the Boston Police Department." *The Review of Economics and Statistics* 91, no. 1 (2009): 163-177.

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.

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.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. "Does the gender composition of scientific committees matter?." *American Economic Review* 107, no. 4 (2017): 1207-38.

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.

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

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, 1957.

Bennett, Dawn, Sally Macarthur, Cat Hope, Talisha Goh, and Sophie Hennekam. "Creating a career as a woman composer: Implications for music in higher education." *British Journal of Music Education* 35, no. 3 (2018): 237-253.

Bennett, Dawn, Sophie Hennekam, Sally Macarthur, Cat Hope, and Talisha Goh. "Hiding gender: How female composers manage gender identity." *Journal of Vocational Behavior* 113 (2019): 20-32.

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.

Bocart, Fabian, Marina Gertsberg, and Rachel AJ Pownall. "Glass ceilings in the art market." *Available at SSRN 3079017* (2018).

Bohnet, Iris, Alexandra Van Geen, and Max Bazerman. "When performance trumps gender bias: Joint vs. separate evaluation." *Management Science* 62, no. 5 (2016): 1225-1234.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. "The dynamics of discrimination: Theory and evidence." American economic review 109, no. 10 (2019): 3395-3436.

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.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. "Stereotypes." *The Quarterly Journal of Economics* 131, no. 4 (2016): 1753-1794.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. "Beliefs about gender." *American Economic Review* 109, no. 3 (2019): 739-73.

Bornmann, Lutz, Rüdiger Mutz, and Hans-Dieter Daniel. "Gender Differences in Grant Peer Review: A Meta-Analysis." *Journal of Informetrics* 1 (2007): 226–238.

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 (2002): 215–226.

Brecher, Ellyn, Jennifer Bragger, and Eugene Kutcher. "The structured interview: Reducing biases toward job applicants with physical disabilities." *Employee Responsibilities and Rights Journal* 18, no. 3 (2006): 155-170.

Breda, Thomas, and Mélina Hillion. "Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France." *Science* 353, no. 6298 (2016): 474-478.

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.

Cannizzo, Fabian, and Catherine Strong. "'Put some balls on that woman': Gendered repertoires of inequality in screen composers' careers." *Gender, Work & Organization* 27, no. 6 (2020): 1346-1360.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. "Are referees and editors in economics gender neutral?." *The Quarterly Journal of Economics* 135, no. 1 (2020): 269-327.

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.

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.

Chari, Anusha, and Paul Goldsmith-Pinkham. "Gender representation in economics across topics and time: Evidence from the NBER summer institute." (2017). Unpublished manuscript.

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.

Diamond, Rebecca, Tim McQuade, and Franklin Qian. "The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco." *American Economic Review* 109, no. 9 (2019): 3365-94.

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.

Dovidio, John F., and Samuel L. Gaertner. "Aversive racism." *Advances in Experimental Social Psychology* 36 (2004): 4–56.

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

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

Fryer Jr, Roland G. "Belief flipping in a dynamic model of statistical discrimination." *Journal of Public Economics* 91, no. 5-6 (2007): 1151-1166.

Fryer, Roland, and Matthew O. Jackson. 2008. "A Categorical Model of Cognition and Biased Decision-Making." *B.E. Journal of Theoretical Economics* 8 (1): Article 1935–1704.

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.

Glover, Dylan, Amanda Pallais, and William Pariente. "Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores." *Quarterly Journal of Economics* 132, no. 3 (2017): 1219-1260.

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.

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

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.

Heckman, James J. "Detecting discrimination." *Journal of Economic Perspectives* 12, no. 2 (1998): 101–116.

Hengel, Erin. "Evidence from peer review that women are held to higher standards." (2017). Unpublished manuscript.

Honigsberg, Colleen, and Matthew Jacob. "Deleting misconduct: The expungement of BrokerCheck records." *Journal of Financial Economics* 139, no. 3 (2021): 800-831.

Hospido, Laura, and Carlos Sanz. "Gender gaps in the evaluation of research: evidence from submissions to economics conferences." *Oxford Bulletin of Economics and Statistics* 83, no. 3 (2021): 590-618.

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

Kempf, Elisabeth, and Margarita Tsoutsoura. "Partisan professionals: Evidence from credit rating analysts." *The Journal of Finance* 76, no. 6 (2021): 2805-2856.

Kessler, Judd B., Corinne Low, and Colin D. Sullivan. "Incentivized resume rating: Eliciting employer preferences without deception." *American Economic Review* 109, no. 11 (2019): 3713-44.

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 (2013): 603–625.

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.

Leonard, Marion. "Constructing histories through material culture: popular music, museums and collecting." *Popular music history* 2, no. 2 (2007).

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. "Expectations of brilliance underlie gender distributions across academic disciplines." *Science* 347, no. 6219 (2015): 262-265.

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 (2013): 231–293.

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

McClary, Susan, and Sam de Boise. "An Interview with Professor Susan McClary: The Development of Research on Gender and Music." *PER MUSI: Revista Academica de Musica* 39 (2019).

McCormick, Lisa. "Higher, faster, louder: Representations of the international music competition." *Cultural Sociology* 3, no. 1 (2009): 5-30.

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

Miller, Diana L. "Gender and the artist archetype: Understanding gender inequality in artistic careers." *Sociology Compass* 10, no. 2 (2016): 119-131.

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* 109, no. 41 (2012): 16474-16479.

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* (2014). https://www.bsomusic.org/stories/by-the-numbers-female-composers/

O'Bannon, Ricky. "What Data Tell us about the 2015–16 Orchestra." *Baltimore Symphony Orchestra* (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.

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 (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.

Petersen, Trond, and Ishak Saporta. "The opportunity structure for discrimination." *American Journal of Sociology* 109, no. 4 (2004): 852-901.

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* 111, no. 12 (2014): 4403-4408.

Robin, William. "What Du Yun's Pulitzer Win Means for Women in Classical Music." *The New Yorker*, April 13, 2017

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

Schwartzstein, Joshua. "Selective attention and learning." *Journal of the European Economic Association* 12, no. 6 (2014): 1423-1452.

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.

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

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/

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.

Swim, Janet K., Kathryn J. Aikin, Wayne S. Hall, and Barbara A. Hunter. "Sexism and Racism: Old-fashioned and Modern Prejudices." *Journal of personality and social psychology* 68, no. 2 (1995): 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.

Tang, Johnny Jiahao. "Individual Heterogeneity and Cultural Attitudes in Credence Goods Provision." *European Economic Review* 126 (2020): 103442.

Ting, Chaowen. “The Data Speak: Women Composer Representation in the 2016–17 U.S. Collegiate Orchestra Repertoire Selection.” *International Alliance of Women in Music 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.

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* 112, no. 17 (2015): 5360–5365.

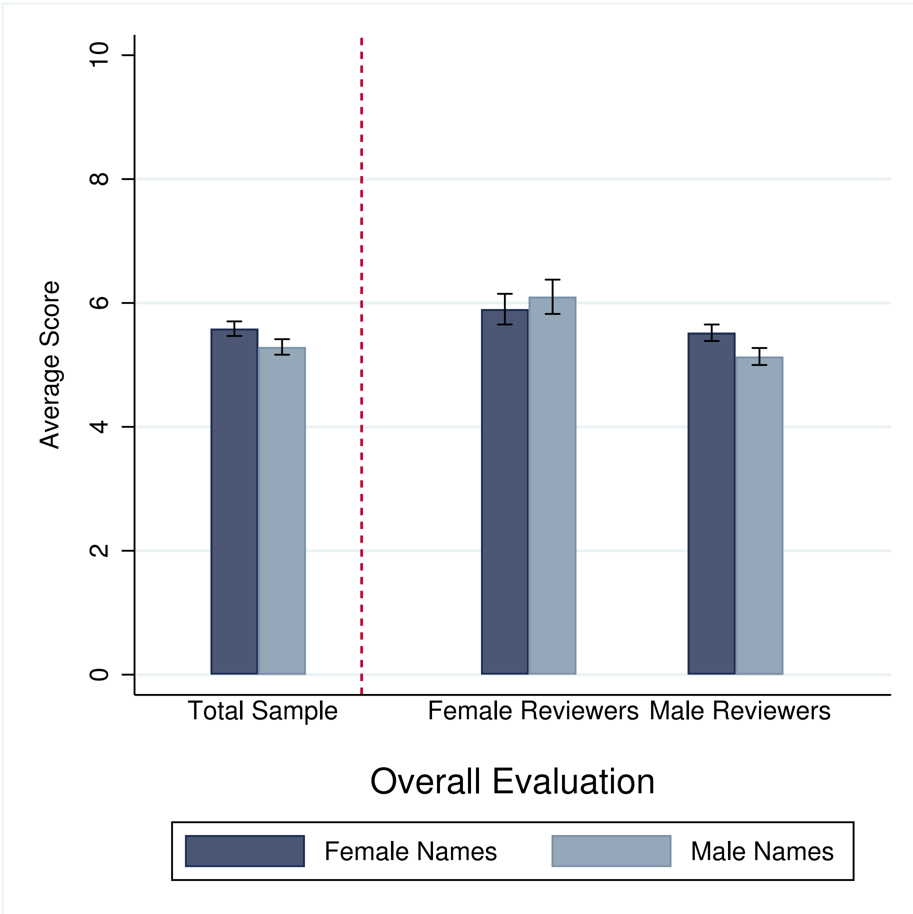
Wu, Alice H. "Gendered language on the economics job market rumors forum." In *AEA Papers and Proceedings*, vol. 108, pp. 175-79. 2018.

Wu, Alice H. "Gender bias among professionals: an identity-based interpretation." *Review of Economics and Statistics* 102, no. 5 (2020): 867-880.

Ye, Junting, and Steven Skiena. "The secret lives of names? name embeddings from social media." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3000-3008. 2019.

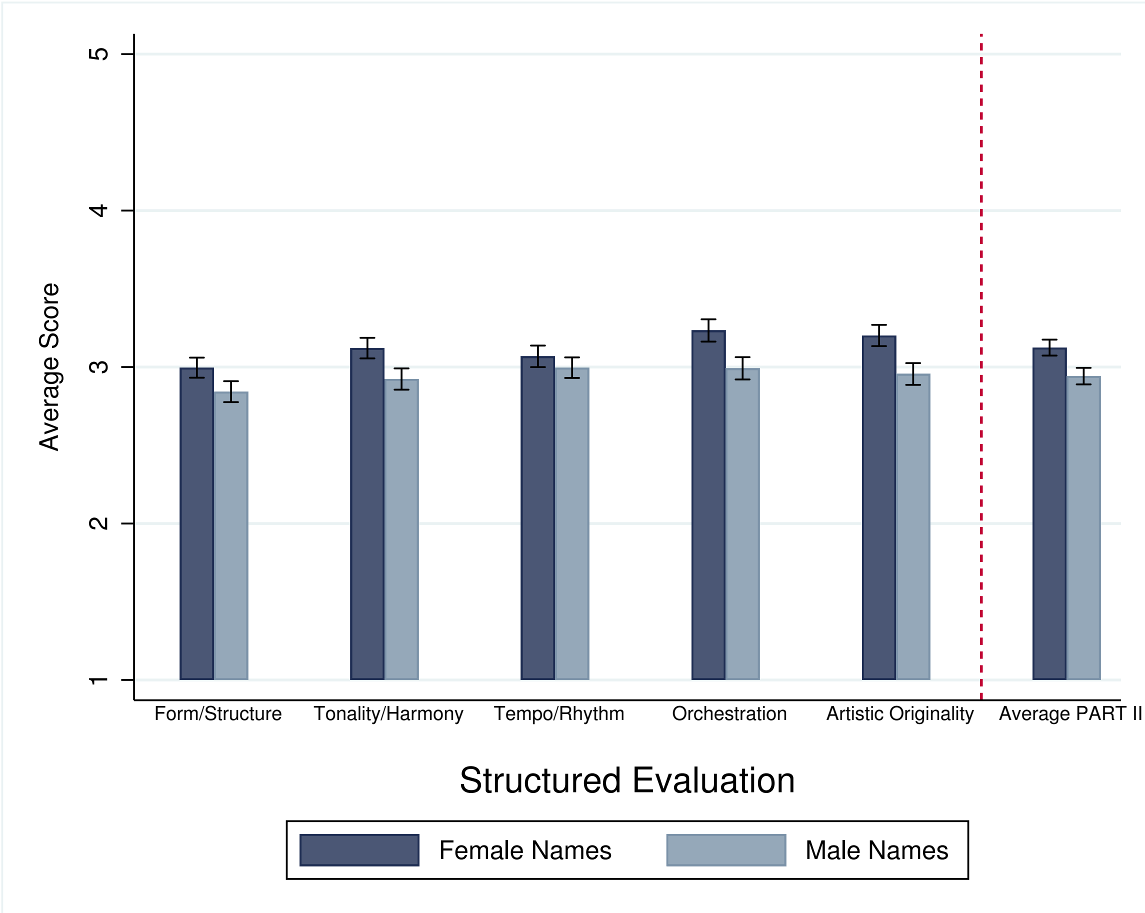
9 Figures and Tables

Figure 1: Graphical analysis of overall evaluation



Note: 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.

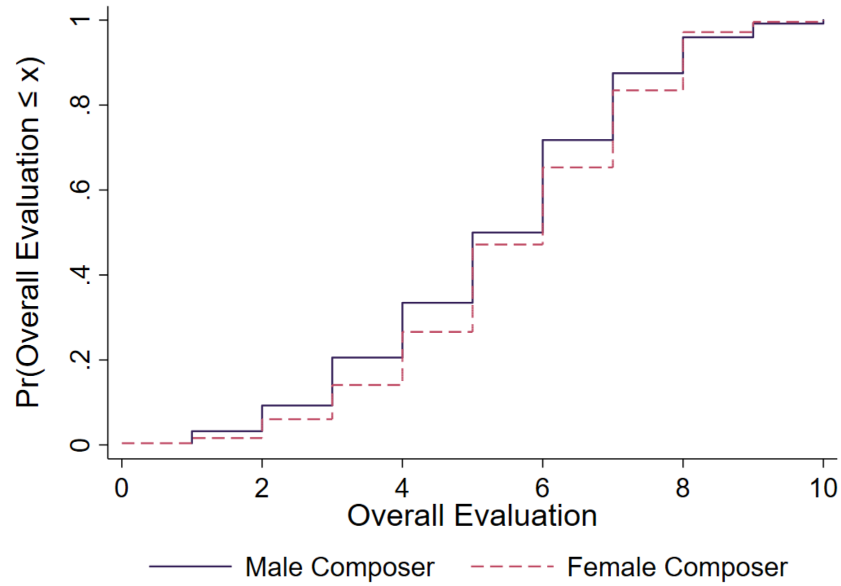
Figure 2: Graphical analysis of structured evaluation



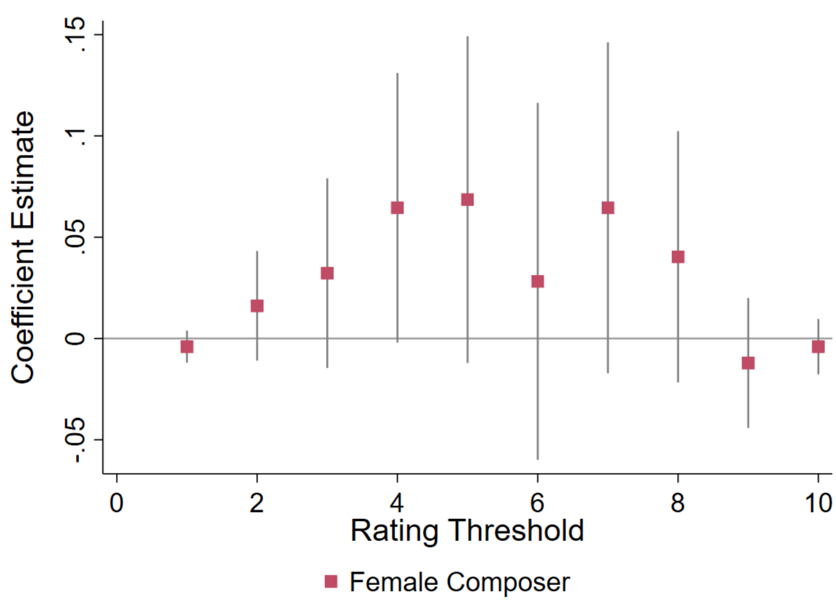
Note: All the bars represent the average scores in Part II structured evaluation (scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong") for compositions associated with female names vs. male names. Compositions are evaluated in five categories.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

Figure 3: Value of Composers' Gender over Selectivity Distribution

(a) Empirical CDF of Overall Evaluation



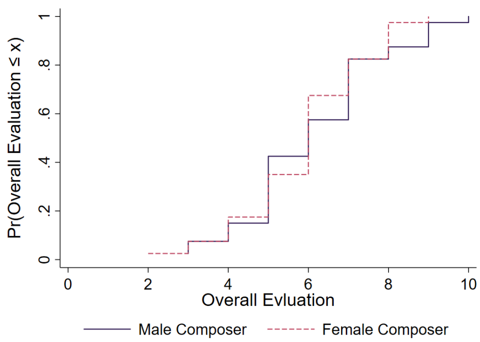
(b) Linear Probability Model for *Female Composers*



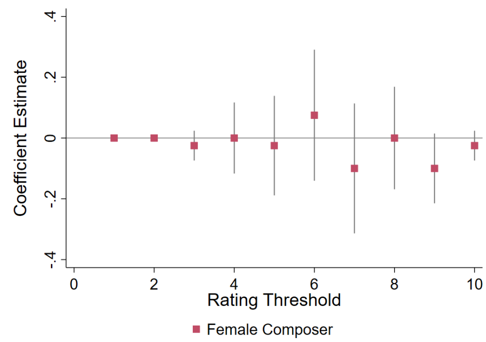
Note: Empirical CDF of *Overall Evaluation Score* is in panel (a). The CDFs show the share of compositions with each characteristic (ex: written by hypothetical male or female composers) with a overall evaluation score less than or equal to each value. The counterfactual callback rates for *female names* is in panel (b), which plot the equivalent "callback rate" for that group. That is, the composition for that group receive the score above the callback rate when it is set at any given rating threshold. Error bars represent 95% confidence intervals. The confidence intervals are calculated from a linear probability model where the dependent variable is an indicator for being at or above a threshold and the independent variable is a dummy variable indicating the composer's characteristic (i.e. female).

Figure 4: Composers' Gender by Reviewers' Gender over Selectivity Distribution

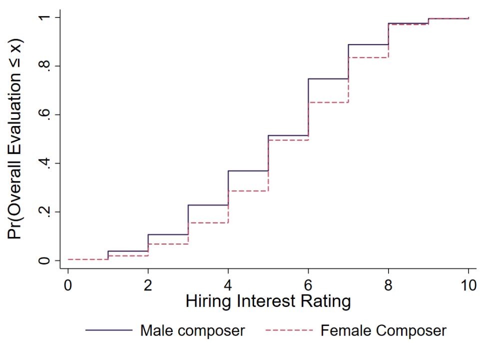
(a) Empirical CDF of Overall Evaluation, Female Reviewers



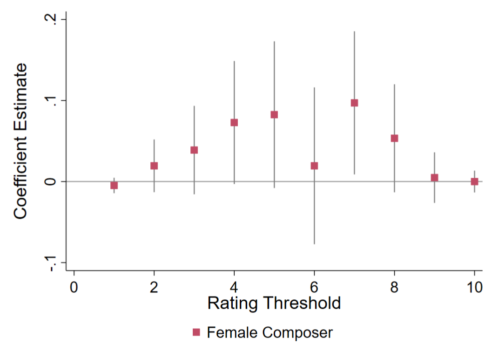
(b) Linear Probability Model for *Female Composers*, Female Reviewers



(c) Empirical CDF, Male Reviewers



(d) Linear Probability Model for *Female Composers*, Male Reviewers



Note: Empirical CDF of *Overall Evaluation Score* is in panel (a) and (c) and the counterfactual callback plot for *female names* is in panel (b) and (d). The top row shows the empirical CDF and counterfactual callback rate when the reviewers are female, while the bottom row shows similar figures when the reviewers are male. The CDFs show the share of compositions with each characteristic (ex: written by hypothetical male or female composers) with a overall evaluation score less than or equal to each value. The counterfactual callback plot shows the equivalent "callback rate" for that demographic group (in this case, female composers). That is, the share of the composers that would be called back if the rating threshold is set at any given value. Error bars represent 95% confidence intervals. The confidence intervals are calculated from a linear probability model where the dependent variable is an indicator for being at or above a threshold and the independent variable is the dummy variable indicating a composer's characteristic (i.e. female).

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: Overall Results for Gender Bias in Music Evaluation

VARIABLES	Overall (1)	Form (2)	Tonality (3)	Tempo (4)	Orch (5)	Artistic (6)	Average score (2)-(6) (7)
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
Standard deviation of outcome variable	1.920	1.036	1.057	1.059	1.128	1.088	0.822
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 names 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 represents the regression coefficient. All regressions are controlled for reviewer and composition fixed effects. 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 3: 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	-0.380***	-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
Standard deviation of outcome variable	1.920	1.036	1.057	1.059	1.128	1.088	0.822
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 the 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 represents the regression coefficient. All regressions are controlled for reviewer and composition fixed effects. 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: 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	-0.278	-0.184	-0.157	-0.301*	-0.0379	-0.0278	-0.141
	(0.311)	(0.168)	(0.205)	(0.147)	(0.176)	(0.286)	(0.115)
Observations	80	80	80	80	80	80	80
R-squared	0.602	0.305	0.437	0.408	0.437	0.400	0.464
Panel B: Male Professor							
Female names	0.402**	0.208*	0.272***	0.161	0.289***	0.293***	0.244***
	(0.162)	(0.118)	(0.100)	(0.119)	(0.109)	(0.107)	(0.0796)
Observations	412	412	412	412	412	412	412
R-squared	0.624	0.419	0.456	0.478	0.450	0.459	0.504
Panel C: Total Sample							
Female names	0.402**	0.209*	0.273***	0.160	0.291***	0.293***	0.245***
	(0.162)	(0.118)	(0.100)	(0.119)	(0.109)	(0.107)	(0.0797)
DID result (Female names X Female Professor)	-0.668*	-0.368*	-0.409*	-0.470**	-0.299	-0.322	-0.374***
	(0.353)	(0.210)	(0.227)	(0.182)	(0.212)	(0.283)	(0.141)
Observations	492	492	492	492	492	492	492
R-squared	0.626	0.403	0.452	0.469	0.454	0.446	0.502

Note: All the results are based on regressions with female names 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. All regressions are controlled for reviewer and composition fixed effects. 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 5: 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	0.0261	0.106	0.191	0.0721	0.181	0.173	0.145*
	(0.185)	(0.146)	(0.122)	(0.132)	(0.140)	(0.128)	(0.0833)
Observations	260	260	260	260	260	260	260
R-squared	0.585	0.364	0.411	0.443	0.450	0.471	0.490
Panel B: Older Professor (≥45)							
Female names	0.536**	0.164	0.199	0.103	0.283**	0.278*	0.205*
	(0.239)	(0.155)	(0.134)	(0.158)	(0.128)	(0.159)	(0.118)
Observations	236	236	236	236	236	236	236
R-squared	0.656	0.427	0.501	0.479	0.453	0.455	0.504
Panel C: Total Sample							
Female names	0.571**	0.188	0.208	0.102	0.300**	0.301*	0.220*
	(0.236)	(0.153)	(0.134)	(0.155)	(0.127)	(0.155)	(0.116)
Panel C: Total Sample							
Female names	0.571**	0.188	0.208	0.102	0.300**	0.301*	0.220*
	(0.236)	(0.153)	(0.134)	(0.155)	(0.127)	(0.155)	(0.116)
DID result (Female names X Younger Professor)	-0.510*	-0.0572	-0.00792	-0.0311	-0.101	-0.105	-0.0606
	(0.304)	(0.214)	(0.180)	(0.204)	(0.189)	(0.205)	(0.145)
Observations	496	496	496	496	496	496	496
R-squared	0.622	0.397	0.447	0.464	0.452	0.447	0.495

Note: All the results are based on regressions with female names 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. All regressions are controlled for reviewer and composition fixed effects. 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 6: 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
	(0.232)	(0.179)	(0.174)	(0.187)	(0.186)	(0.164)	(0.122)
Observations	168	168	168	168	168	168	168
R-squared	0.626	0.365	0.378	0.438	0.431	0.380	0.460
Panel B: Assistant Professor							
Female names	0.406	0.387*	0.0793	0.286	0.242*	0.387**	0.276**
	(0.239)	(0.217)	(0.143)	(0.172)	(0.126)	(0.185)	(0.111)
Observations	116	116	116	116	116	116	116
R-squared	0.671	0.440	0.524	0.447	0.593	0.547	0.588
Panel C: Associate Professor							
Female names	0.425	-0.0458	0.196	-0.262	0.458*	0.162	0.102
	(0.359)	(0.191)	(0.248)	(0.190)	(0.246)	(0.233)	(0.138)
Observations	88	88	88	88	88	88	88
R-squared	0.634	0.385	0.418	0.547	0.527	0.471	0.521
Panel D: Full Professor							
Female names	0.852**	0.507**	0.430**	0.356	0.205	0.580**	0.415**
	(0.360)	(0.222)	(0.197)	(0.265)	(0.199)	(0.244)	(0.188)
Observations	108	108	108	108	108	108	108
R-squared	0.638	0.501	0.546	0.497	0.439	0.499	0.524
Panel E: Total Sample							
Female names	0.0915	0.102	0.128	0.0956	0.209*	0.153	0.138
	(0.167)	(0.135)	(0.111)	(0.128)	(0.122)	(0.118)	(0.0832)
DID result (Female names X Tenured Professor)	0.540*	0.142	0.190	-0.0243	0.0963	0.237	0.128
	(0.308)	(0.208)	(0.190)	(0.217)	(0.199)	(0.208)	(0.149)
Observations	496	496	496	496	496	496	496
R-squared	0.622	0.398	0.449	0.464	0.452	0.450	0.496

Note: All the results are based on regressions with female names 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. All regressions are controlled for reviewer and composition fixed effects. 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 7: Robustness Check: Dropping Those Whose Response Times are Short

VARIABLES	Overall	Form	Tonality	Tempo	Orch	Artistic	Average score (2)-(6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female names	0.284*	0.152	0.221**	0.0753	0.244**	0.221**	0.183***
	(0.148)	(0.106)	(0.0918)	(0.100)	(0.0991)	(0.0969)	(0.0697)
Average score of outcome variable	5.438	2.919	3.022	3.032	3.113	3.079	3.033
Standard deviation of outcome variable	1.920	1.036	1.057	1.059	1.128	1.088	0.822
Observations	472	472	472	472	472	472	472
R-squared	0.619	0.397	0.453	0.476	0.454	0.452	0.504

Note: All the results are based on regressions with female names 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. Those whose response times are less than 10% of the bottom tail are dropped. The top row represent the regression coefficient. All regressions are controlled for reviewer and composition fixed effects. 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 8: Robustness Check: Selection into the Survey

	In sample	Out of sample	Difference between those in sample and out sample
Female (0/1)	0.1612	0.1799	-0.019
	(0.0331)	(0.0124)	
Public School (0/1)	0.6363	0.6165	0.0198
	(0.0439)	(0.0157)	
Tenured (0/1)	0.4117	0.6842	-0.272***
	(0.0453)	(0.0164)	

Note: Those can we cannot identify the rank or the visiting professors are not included in this test. There are 167 out 1081 faculty members who we cannot identify the rank. 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: 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
Standard deviation of outcome variable	1.920	1.036	1.057	1.059	1.128	1.088	0.822
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.

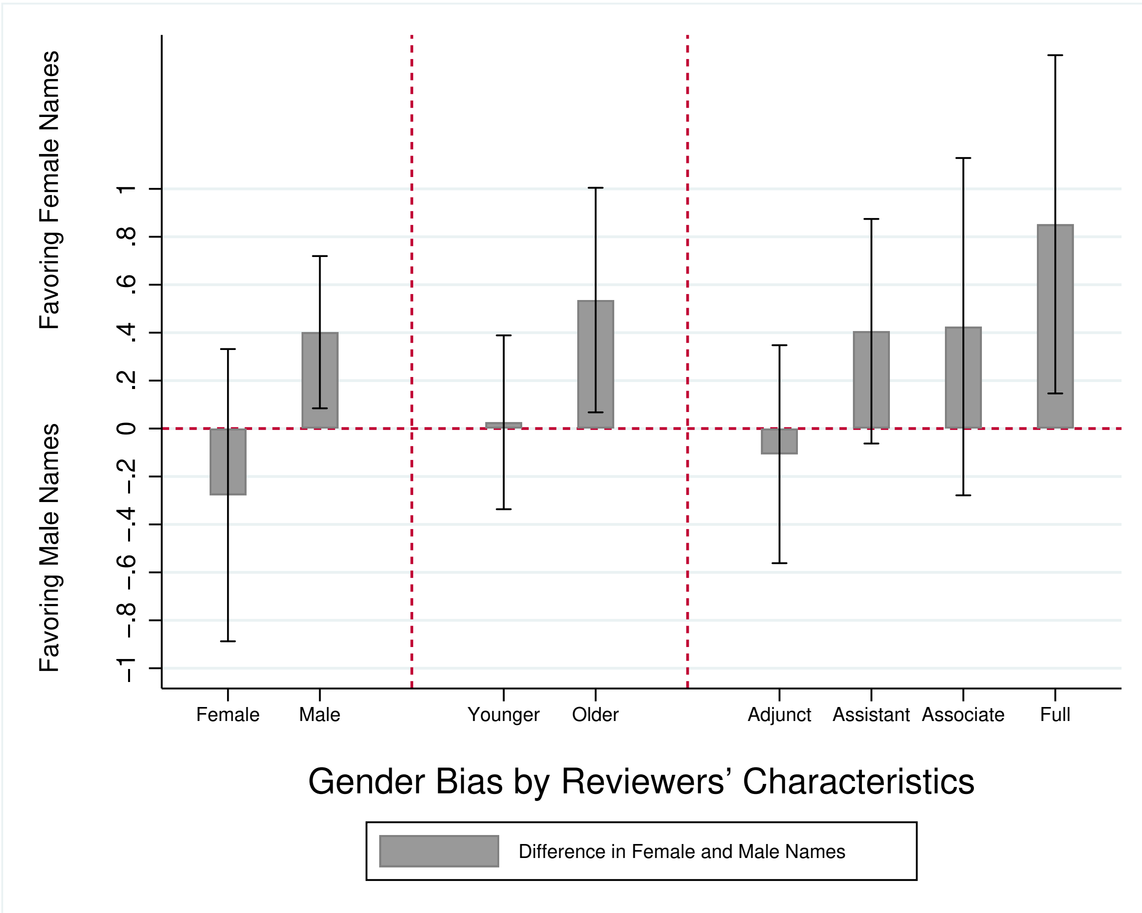
Table 10: Robustness Check: Dropping Outliers (In-group Bias)

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
Standard deviation of outcome variable	1.920	1.036	1.057	1.059	1.128	1.088	0.822
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.

10 Appendix A: Figures and Tables

Figure A-1: Heterogeneity in gender bias



Note: All the bars represent the gender bias, calculated as the average score difference between compositions with female names and compositions with male names. Positive (negative) numbers indicate higher (lower) scores 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 manuscript and Tables 5-7). Error bars represent 95% confidence intervals.

Figure A-2: One Example Page from The Score

11 Appendix B: Survey Materials

11.1 Invitation Letter

Dear Composition Faculty,

I would like to invite you to participate in a research study of composition evaluation conducted at xxx (anonymous institution). The purpose of this study is to better understand criterion used in assessing musical compositions. You will be asked to evaluate four 10-minute orchestral pieces, and the survey will take approximately 40-60 minutes to complete. To compensate for your time, we will provide a seventy-five dollars ($75) e-gift card upon completion of the survey.

The records of this study will be kept private, and your personal information will not be identifiable in any future reporting of results. Research records will be kept in password-protected files; only the researchers will have access to the records. The risks are no more than experienced in everyday life while using the internet. Study records will be kept confidential to the extent required by law.

To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB may review study records. The Office of Human Research Protections may also look at study records. If you have any questions about the study, you may contact Dr. xxx at telephone (xxx) xxx-xxxx, or xxx xxx.xxxx.edu.

If you have any questions about your rights as a research subject, you may contact Ms. xxx, xxx (anonymous institution) at (xxx) xxx-xxxx. Your participation in this study is voluntary. You do not have to be in this study if you don't want to be. You have the right to change your mind and leave the study at any time without giving any reason and without penalty. You will be given a copy of this consent form to keep. You do not waive any of your legal rights by agreeing to be in the study. Your completion of this survey provides your consent to participation.

Thank you for participating in this survey.

Please follow this link to the Survey:

Take the Survey

Or copy and paste the URL below into your internet browser:
https://xxxxx

Follow the link to opt out of future emails:
Click here to unsubscribe

(Email Signature)

11.2 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)	o	o	o	o	o
Tonality / Harmony (2)	o	o	o	o	o
Tempo / Rhythm (3)	o	o	o	o	o
Orchestration (4)	o	o	o	o	o
Artistic Originality (5)	o	o	o	o	o

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