Gender Bias in Competitive Music Composition Evaluation: An Experimental Study

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

August 23, 2023

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 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: Institute of Economics, Academia Sinica, Taipei, 11529 (email: yating@econ.sinica.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, National Tsing Hua University, Academia Sinica, 2021 Asia Impact Evaluation Conference, 2023 NTU-AS Young Economist Workshop, and Virtual East Asia Experimental and Behavioral Economics Seminar Series.

1. Introduction

The underrepresentation of women in various academic fields remains a pressing global concern. Much of the discussion on this issue has focused on STEM fields, yet women are also underrepresented in non-STEM fields, such as economics and philosophy. As of the mid-2000s, women accounted for less than 35 percent of PhD students and 30 percent of assistant professors in economics (Lundberg and Stearns 2019). This gender imbalance has sparked lively discussion about the disparate treatment of women and gender-biased atmosphere in the economics profession (Wu 2018; Wu 2020) and instigated new research on gender bias published in economics journals and presented at economics conferences (Hengel, 2018; Hospido and Sanz, 2019; Chari and Goldsmith-Pinkham, 2017; Card et al., 2020). In 2019, the *Journal of Economic Perspective* hosted a special symposium on women in economics (Bayer and Rouse, 2016; Avilova and Goldin, 2018; Boustan and Langan, 2019; Lundberg and Stearns, 2019; Buckles, 2019). A prevailing argument from this symposium contended that female candidates face unequal evaluations during selection processes, encompassing both interviews and peer reviews. This inequitable treatment often stems from discriminatory biases against minority groups (Greenwald and Banaji, 1995; Greenwald and Krieger, 2006; Bayer and Rouse, 2016).[1] Another contributing factor is "in-group bias," where individuals show favoritism towards members of their own group. For example, in male-dominated field, men might show favoritism towards other men (Tajfel et al., 1971; Bernhard et al., 2006; Chen and Li, 2009; Goette et al., 2012; Sandberg, 2018). Moreover, Dustan et al. (2022) highlight the concept of women's second-order belief as a potential explanation for this gender imbalance. Specifically, this belief suggest that women believe that leaders in STEM fields

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

believe that women are bad at doing science. Such beliefs about beliefs could steer women away from STEM fields (Dustan et al. 2022). These factors combine to form barriers for women, limiting their opportunities for career advancement.

Given this background, we believe that a study of gender representation in professional music composition would shed light on the origins of gender representation disparities in a non-STEM field. While the traditional STEM fields display significant gender imbalance, the field of music composition manifest an even more pronounced underrepresentation of women, surpassing other traditionally male-dominated disciplines. According to Leslie et al. (2015), only about 16 percent of U.S. doctorates in music composition were awarded to women. This was the lowest percentage among all fields, with rates even lower than those in physics (18%), computer science (19%), and economics (35%). This gender disparity is even more stark in the professional realm. In 2014, women occupied only 15 percent of the U.S. composition faculty positions, and only 9 percent of the prestigious composition prizes were awarded to women (O'Bannon, 2014). A study revealed that only 2% of the pieces performed by symphony orchestras worldwide in 2018 – 2019 were written by women (Brown, 2018).

Beyond the statistics, research has highlighted the pervasive gender discrimination faced by women in the music composition profession, such that female composers often hide their gender identities to mitigate unfavorable outcomes (Bennett et al. 2018, Bennett et al. 2019, Cannizzo and Strong 2020). Gender bias in music evaluations (Goldin & Rouse, 2000; Leonard, 2007; Miller 2016), which has been well-documented in the literature, is certainly a component of gender discrimination. Despite this, we have witnessed more conversations about gender and diversity issues in composition and some small steps have been made (Robin, 2017; McClary and de Boise

2019). For example, the number of female composers who have joined prestigious programs in recent years has increased. In a historic milestone for 2017, all three Pulitzer Prize for Music finalists were women. A social movement to promote diversity in music is also growing via social media, with the hashtag #HearAllComposers gaining traction.

We designed an experiment to examine whether college composition faculty members evaluate compositions fairly when the music is randomly associated with a gendered name, and whether the evaluations differ with more structured evaluation methods, such as giving detailed ratings on various musical dimensions.[2] 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 from randomly selected 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 participants have a training background in composition. We randomly assigned participants to one of the two groups, with randomization done at the institutional level, meaning that 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 were newly composed, with duration ranging from 8 to 10 minutes. They were commissioned by the Georgia Tech School of Music, with all pieces being selected based on the same criteria. Each piece was composed by a different professional composer with a doctoral

[2] Our experimental design departs from the conventional "resume audit experiment" paradigm, primarily due to the minimal deception involved. Notably, our compositions were crafted by real composers, ensuring the authenticity and avoiding the use of fabricated materials typically employed in standard audit experiments. Moreover, we did not measure the conventional "callback" in a traditional hiring experiment, as it can introduce costliness and potential bias, leading to preferential selection of candidates with higher acceptance likelihood. Rather, our evaluation approach bears some similar merits as in the Incentivized Resume Rating methodology (IRR) proposed by Kessler et al. (2019), which can be used to elicit employer preferences without deception to detect discrimination. For an in-depth exploration of our methodology, please refer to the corresponding methodology section.

degree in composition. In the evaluation, we provided both the music scores and the audio recordings from live performances conducted by the same conductor and performed by university orchestras to ensure consistent recording and performance quality. Our experimental design replicates the setting of real-world competitions.

Our results indicate that compositions associated with female names receive higher ratings than those linked to male names. This finding is unexpected based on our initial hypothesis and remains consistent in more structured evaluations when participants assess compositions across various subcategories. We also observe some suggested evidence of heterogeneity: Male evaluators tend to favor compositions by female composers, while female evaluators show no discernible gender bias. Senior faculty members display a stronger preference for female composers compared to their junior counterparts. These results align with other 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 tests in Israel (Bar and Zussman, 2020).

The present paper also speaks to the literature on discrimination. A significant portion of this literature documents evidence of discrimination against racial, ethnic, and gender minority groups (for a review of field experiments on discrimination, see Bertrand and Duflo, 2016). The primary economic models addressing discrimination are Becker's taste-based model (Becker, 1957; Charles and Guryan, 2008) and the statistical discrimination model (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977). Becker's model suggest that in the labor market, some individuals have a non-pecuniary aversion to hiring members from minority groups. Conversely, the statistical

[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.

discrimination model is based on a rational perception of skill distribution among groups, predicting that employers or evaluators would favor majority groups (e.g., white and male) due to their perceived higher productivity. Another strand in the literature centers on the belief-based model (Fryer 2007; Bohren et al. 2019). This model describes 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 disappears or even reverses. Our experiment examines gender bias in a unique field characterized by a significant gender imbalance with the identities of both evaluators and evaluatees. Our findings suggest a reversal of gender bias towards female favoritism, aligning more with the belief-based model. Notably, we observe a pattern where senior faculty members display a pronounced preference for women. This pattern could further support a key premise of the belief-based model in a dynamic setting: the presence of individuals who recognize biases in others. Given senior faculty members' tenure in the discipline, it is more likely that they recognize such biases.

This reversal is echoed in other contexts. For example, female artists, despite being less likely to transition from primary to secondary art markets, attain a significant premium on their sold artworks once they do (Bocart et al. 2022). In academia, Mengel et al. (2019) find that while junior female instructors receive lower teaching evaluations than male peers for similar courses, senior female instructors outperform male instructors in evaluations. In a large Q&A online forum, Bohren et al. (2019) find discrimination against newly created women accounts. Once these accounts achieve advanced status, these biases reverse—advanced female accounts earn more positive votes than their male counterparts at a similar level. More recently, Card et al. (forthcoming) recognizes a similar pattern, finding that the Econometric Society admitted more female than male fellows during the 2010-2018 period, reversing the historical trend to now favor female scholars. They explain that fellows may exhibit positive values demonstrating more equal gender representation to showcase their values to the academic society. However, the results are less consistent with the taste-based model and statistical model in a dynamic setting. This is because such reversal biases favoring women would not be expected in a traditionally male-dominated field.

In-group bias is another potentially significant explanation for observed gender discrimination. If in-group bias exists in evaluation, the dominant group (e.g. males) in a particular discipline will perpetuate unbalanced representation. Our experiment, one that mimics a real-world music competition, would add to the literature on in-group bias. Most evidence of in-group bias has come from experiments in laboratory settings, which have shown that in-group favoritism will arise in both natural social groupings and artificially 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 challenging to test to what extent this favoritism persists in real-world settings. Studies have, so far, found mixed results. In the professional sports setting, for instance, scholars have found discrimination toward 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 in juvenile court cases. 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). Gender in-group bias has been found in teaching evaluations (Boring 2017). Yet some research suggests no gender in-group bias in other contexts, such as journal reviews (Abrevaya and Hamermesh, 2012) and grading (Feld et al., 2016), and even out-group bias in Bar and Zussman's (2020) study of driving tests. Overall, the existence and extent of in-group bias on gender, race, and nationality is not conclusive and demands more empirics.

Our experiment crafts a "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 circumvents these issues.

One primary challenge to testing in-group bias in real-world settings is the often unavailability of evaluator's identity. In many experimental studies that testing discrimination, researchers can typically only record outcomes, such as the callback rate in Bertrand and Mullainathan (2004) (For a more comprehensive review on field experiments testing discrimination, see Bertrand and Duflo (2017)). Similarly, Goldin and Rouse's (2000) study on orchestra audition outcomes—a context similar to ours—the researchers lacked information on the juries' identities. To address this challenge, we specifically recruited faculty reviewers, ensuring both male and female evaluators with consistent demographics across experimental groups.

Another concern is the potential correlation between an evaluators' group identity, such as gender, and candidates' quality. This non-random matching of evaluators to candidates makes it difficult to pinpoint in-group bias, evident in observational studies, including journal review (Abrevaya and Hamermesh, 2012), tenure review (Bagues and Esteve-Volart, 2010), and professional sports competition (Price and Wolfers, 2010; Sandberg, 2018). In our experiment, all faculty reviewers were given to four compositions with randomly assigned gendered names (two female and two male names). This randomization design circumvents identification challenges.

A third challenge arises when a group's identity, such as race and gender, is not exogenous. This could lead to overstated discriminatory effects in market outcome due to other unobserved variables (Siegelman and Heckman, 1993; Heckman, 1998). In many non-laboratory studies, the actual productivity or skill of a candidate is rarely observed directly and can be confounded with minority identities, like race and gender. Our approach uses the same composition but assigns both male and female names to it. Specifically, composition #1 is assigned with a male name in randomization group A and with a female name in group B. This method ensures that the unobservable characteristics cancel out to zero for the same audit pair.

A fourth difficulty in testing in-group bias comes from endogenous behaviors in typical hiring and evaluation studies. In hiring studies, for example, interviewees' behaviors may be endogenous to the recruiter's race or gender. In our field experiment, we only rely on music scores and recordings rather than actual people, and are thus able to eliminate this endogenous behavioral factor, providing a clearer analysis of gender or in-group bias.

There are several implications to our findings. First, in-group bias may be malleable and can vary based on the real-world context, diverging from patterns observed in laboratory settings. Second, even in a highly professional setting where gender imbalance persists, evaluators can update their beliefs and potentially reverse gender bias upon observing the commendable or equal performance of the minority group. Finally, biases in treating a 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).

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

Composition competitions are important events for composers' career advancement. These competitions are the primary vehicles used to identify young talents ready to launch international careers (McCormick, 2009). The competitions are hosted frequently 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 develop potential employment opportunities. Composers' works are selected for performance through such competitions, and these performances often lead to future engagements and even commissions (Whitacre, 2009; Watts, 2018; Murphy, 2012; Doolittle, 2018). Research shows that prior success in music competitions has strong implications for later success. 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 was predictive of subsequent market success.

Our experimental design follows the form of hosting a formal composition competition. The organizer sends out invitations to professional composers to participate as reviewers. The scores, often along with a recording, are uploaded to a digital platform and the reviewers give evaluations of the compositions. Although professional composers are able to judge the quality of a composition based solely on the music score, it is common for competition to provide recordings to facilitate the evaluation process. We adopted this same design in this study. One of the authors, a professional conductor, conducted and recorded all four pieces of music to ensure the quality of recordings are commensurate with professional standards. We could have created a control group without any composer names as the baseline, yet such anonymous design deviates from common practices in the profession. Esteemed and highly influential composition competitions, such as Guggenheim, Fromm, Koussevitzky, ASCAP, Barlow, are not anonymous, and we follow suit. Because our study focuses on a group of highly specialized professionals, it is distinctly challenging to find qualified subjects. We sought to maximize the results' statistical power by not adding another control group. Since our main purpose is to understand the relative level of gender bias and not the absolute level of bias, we keep with the standard practice of including composers' names.

3. Experimental Design

In this section, we present our experimental design to measure gender bias in the professional music composition setting. In this experiment we also aim to test the effects, if any, of same-gender in-group bias as found in the experimental literature. Some research suggests that structured evaluation methods—ones that with clear rubrics—could reduce bias, and we test this claim in the experiment (Bragger et al., 2002; Brecher et al., 2006; Levashina et al., 2013).[4] Our hypotheses were:

[4] This study was not pre-registered to ensure that the unique faculty population, who are familiar with google search, did not become aware that they were part of a gender-bias study. In addition, it was not a prevalent requirement from journals at the time of our field experiment. Nonetheless, our hypotheses are clearly defined and straightforward, and the heterogeneous analysis are based on commonly tested covariates and are consistent with the approved survey tools that we submitted to IRB. The primary deviation was our anticipation of a gender bias toward male composers. Consequently, we delved deeper into the ingroup hypothesis to shed light on this unexpected outcome.

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

Hypothesis B: Reviewers would favor compositions associated with a name of their own gender, showing in-group gender bias (favoritism/in-group bias hypothesis).

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

Experiment in Music Composition

Our main objective is to examine gender bias in music competitions by presenting compositions randomly assigned to male and female composer names. All compositions last between 8 and 10 minutes when performed and were composed by four different professional composers, all with a terminal doctoral degree in composition. All the works were chosen by the same conductor using the same selection criteria, and they were all commissioned works by the Georgia Tech School of Music. The compositions had only been performed once in a regional concert in the southwestern U.S., and the reviewers were unfamiliar with the pieces of music before our experiment. This is a common practice in competitions, wherein composers submit recordings of performances 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. All the recordings were made during two performances in the same concert hall conducted by the same conductor in a two-month span. Systemic forces upstream might lead male composers to produce higher-quality compositions. For example, given the prevalence and established presence of men in the field, male composers could have access to superior equipment, enhancing recording quality.

An attractive property of our experimental design circumvents these systemic biases. 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 is the standard practice in composition competitions. We pre-tested the survey with faculty members who have served as composition competition judges. They agreed that this design is 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. Because this population is a highly white-dominated discipline, none of the respondents expressed any comments regarding the race of the composers' names. We tested the names using NamePrism (Ye et al., 2017), a non-commercial ethnicity/nationality classification tool often used in economics studies (Diamond et al., 2019; Tang, 2020; Honigsberg and Jacob, 2021; Kempf and Tsoutsoura, 2021). Our pseudo names are unambiguously (around 90%) classified as white names.

Some observers may be concerned that the actual gender of the composers affects the "gendered" features of the compositions, yet we believe that this is not a significant issue in our study. The compositions in our sample are created by two females and two males, and we assign names to all the compositions. According to the professional musicians, classical symphony compositions usually do not have easily identifiable gendered features, meaning that it is not easy to determine the gender of the composers based on the music alone. Edvenson's (2017) study, for instance, found that listeners could not accurately guess the gender of a composer based solely on listening to the classical music performed. Even if our compositions did exhibit gendered features, our fixed effects empirical design would control for individual composition factors. We are also able to compare results between the compositions created by females and those composed by males.

Recruitment

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 choose 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 was comprised of 1,060 eligible faculty members. Faculty reviewers were randomly assigned to group A or B to evaluate four compositions as if they were participating in a composition competition. Randomization was at the institution level to ensure that faculty members at each institution received compositions associated with the same female and male composer names, preventing sample contamination. We cross-checked for duplicate faculty names in case anyone were transitioning from one institution to another or if certain faculty were affiliated with more than two institutions, ensuring that each faculty member is invited only once per 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 the Qualtrics system through a personalized survey link sent in the invitation email. This study was approved by the Institutional Review Board (IRB) at Georgia Institute of Technology and the subjects were informed of the approval in the email invitation (Appendix B). The subjects would give consent when they filled out the online survey. The reviewers were rewarded with a $75 gift card upon completion of the evaluation. Only 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 was comprised of 124 faculty reviewers (20 female, 103 male, 1 other). The response rate is 11.7%, which is normal for a highly selected group. Research 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). 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 difference in the attrition rate. The sample may still differ in other unobservable dimensions. Nonetheless, the music professionals in our sample 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 introduced the survey to participants as a music evaluation study. In the invitation letter, we emphasized our interest in understanding their criteria for evaluating musical compositions, all the while keeping them unaware of the study's underlying investigation into potential gender bias. Reviewers were instructed to judge the compositions as though they were judging a real competition. The names of the composers were only disclosed within the score and the final question, where reviewers were asked to recommend a winner. This setup aims to mitigate the experimenter effect, an effect that may cause human subjects to change their behavior if they realize their participation in an experimental context. Simultaneously, we can ensure genuine responses from the reviewers without resorting to intentional deception. Our approach significantly diverges from the deceptive practice, and aligns with the acceptable circumstances of non-disclosure proposed by Cooper (2014). First, our approach does not cause more harm to subjects than a deception-free economic experiment would. Second, conducting this survey without our setup would be challenging since concealing the study’s underlying purpose is arguably the best method to uncover implicit bias in responses. Accordig to Charness et al. (2022)’s extensive survey of experimental economists’ attitudes towards deception, our approach—making participants unaware of the experiment—lacks many potential alternatives and is among the least deceptive techniques in "gray area" experimental scenarios that omit information or mislead without outright lying (Charness et al. 2022).[5]

Alternative methods to elicit preference without deception, such as the one proposed by Kessler et al., (2019) where employers knowingly evaluate fictitious resumes, could have been employed. However, our chosen approach bypasses common challenges inherent to traditional audit studies and maintains comparable merits. In resume studies, the concern arises over potential differences between fabricated resumes and real resumes. This concern does not exist in our study, as all the compositions are genuine works written by actual composers. The composers' real names are the only element we altered, serving to protect their true identities. Another prevalent issue in resume studies is the "callback rate," where employers' assessments of resumes could be conflated with their expectations of candidates' acceptance. Our study avoids this problem by asking evaluators to assign scores from 1-10 to all compositions, rather than tracking phone calls to discern evaluators' binary decisions. This approach mirrors the strategy in Kessler et al. (2019), where they assess employers' hiring interest using a Likert scale from 1 to 10, without measuring costly actual callback decisions.

The study's survey contained four parts: general evaluation, structured evaluation, winner recommendation, and demographic information. However, in conventional practice, reviewers typically provide only the overall evaluation as their judgement.[6] Consequently, the interpretation

[5] Our approach arguably involves even less deception than the "gray area" scenario described in Charness et al. (2022)'s study. The most comparable "gray area" scenario is the "Unknown/unpaid participation." The text in this scenario is described as "*The experimenter conducts a field experiment that encourages people to put forth (unpaid) effort or take action, but does not inform the participants that they are in an experiment.*"

[6] Concerns about inducing an "experimenter effect" due to extra questions may arise. However, we believe that it is unlikely because we clearly explained in the beginning of the survey that our goal is to better understand criteria used in assessing musical compositions, making those extra questions seem legitimate. Additionally, we minimized the gender cues presented in the study, and our population has limited knowledge of experimental studies, reducing the probability of priming them about gender cues and their direct effect on the gender bias we aim to test.

of gender bias should largely depend on this score. The overall evaluation was scored on a 0–10 scale.

Since the overall evaluation assessment is highly subjective, we added a structured evaluation in Part II so that reviewers were provided with a structured evaluation format to appraise compositions in more detail. This structured approach aims to eliminate arbitrary bias, including inattention during quick decision-making, or potential sequential spillover effects arising after evaluating traditionally stronger candidates, such as white males (Kessler et al. 2022). By prompting reviewers to examine the works more closely, the structured evaluation can help mitigate the influence of arbitrary bias during the judgment of compositions. The 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 because they consist of the basic elements of musical composition and were suggested by professional composers who have judged composition competitions. The five categories covered both the craft and creativity of music composition and were gender neutral terms. To ensure the integrity of Part II's 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 from Part I once they advanced to Part II.[7] The structured evaluation in Part II was based on a 1–5 Likert scale labeled "Extremely Weak," "Somewhat Weak," "Neutral," "Somewhat Strong," and "Extremely Strong." The scale was pre-tested among faculty who have judged music and they agreed that this chosen scale is commonly used. The survey questions are in Appendix B.

[7] 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.

4. Estimation Strategy

In the summary statistics and the mean evaluation tests (Table 1), we used the original 124 faculty reviewers as the unit of analysis to show the overall descriptive characteristics in the data. To empirically test for gender bias, we treated the unique combination of composition and reviewer as the unit of analysis. The total observations expanded to *N* = 496. Each reviewer evaluated four compositions and the standard errors may be correlated within the same reviewers, so we followed the standard practice of clustering our standard errors by reviewer. Our estimation equation is:

$$\mathrm{Score_{ic}} = \beta_1 \mathrm{Female_{ic}} + \delta_\mathrm{c} + \gamma_\mathrm{i} + \varepsilon_\mathrm{ic} \quad (1)$$

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; β_1 is the coefficient for gender bias. δ_c controls for composition fixed effects; γ_i controls for reviewer fixed effects. This model controls for differences among compositions, so that we compare reviewers' evaluations within the same composition assigned with a female name and a male name. Controlling for reviewer fixed effects should eliminate confounding issues arising from unobservable differences between reviewers. We expected $\beta_1 < 0$ if there is implicit gender bias in favor of male composers relative to female composers. We ran a 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 the same 7 sets of regressions by gender, age, and rank to examine heterogeneity in gender bias. Since the scoring of the compositions is done by the same reviewers and 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 that indicate the same gender as their own, as opposed to compositions assigned with opposite gendered names. Similar to the above equation, we estimate:

$$Score_{ic} = \beta_2 Same\_gender_{ic} + \delta_c + \gamma_i + e_{ic} \qquad (2)$$

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), β_2 should be larger than zero. We also controlled for several fixed effects, including composition fixed effects (δ_c) and reviewer fixed effects (γ_i). Similar to the previous specification, we clustered the standard errors by reviewers.

5. Results

Based on ratings of the same compositions with male and female names, we estimated gender bias through standard regression analysis, clustering standard errors. We employed additional controls for variables that might bias our results. We describe below the analysis's detailed results.

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, accounting for approximately 16% of the overall standard deviation. For comparison, in a separate 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). When we ask the reviewers to pick winners, the compositions with female names were 1.6 times more likely to be selected the winners than were the same compositions associated with male names.

Differences were found in all five categories in the structured evaluation, with female composers receiving an average of 0.188 more points than did male composers on a 1-5 scale (See Figure 2 and Table 2). The precise 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 tested whether reviewers exhibit in-group bias toward the same gender. Contrary to the literature (Nosek et al., 2002; Dovidio and Gaertner, 2004; Moss-Racusin, 2012) and the majority of past laboratory findings, our results show no evidence of in-group bias based on gender—reviewers did not give higher scores to composers of their own gender. In fact, 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 compositions produced by members of the opposite gender. The negative coefficient (-0.38) is nearly 20% of the overall standard deviation. This magnitude is a drastic difference from what is reported in current literature. Sandberg (2018) has, for instance, found that nationalistic in-group bias in the Olympic game is between +7.2% and +23.8% of the overall standard deviation, while our estimated coefficient is in the opposite direction with similar magnitude.

This result is mostly driven by the prevalence of male reviewers favoring 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—0.402 points corresponds to approximately 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 the actual 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).

We note 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. Effects across the Distribution of Overall Evaluation

We follow Kessler et. al (2019) in implementing a counterfactual callback exercise. At each overall evaluation scoring level, we compute an analogous counterfactual callback rate, treating each scoring level as the callback threshold. Under this framework, composers receive a "callback" when their overall evaluation score exceeds that given score, as if in a hiring context. Even though our study is not a hiring experiment, our granular measure—overall evaluation score—enables us to examine potential variations in evaluators' preferences in the tails of the distribution. Unlike typical audit studies, our measure mitigates the disadvantages associated with the binary callback rate variable's sensitivity to a low callback rate environment (Kessler et al. 2019).

Two graphs are presented using this counterfactual callback rate approach. The first graph displays the empirical cumulative distribution function (CDF) of the overall evaluation scores. In our study, we compare the CDFs of female composers and those of male composers. The second graph generates counterfactual callback rates by assuming that evaluators would call back the composers if their overall evaluation scores were at or above a given 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 male and female composers (see Figure 3 panel a). We also show the differences in the counterfactual callback rates (Figure 3 panel b). We observe 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 qualitatively consistent across nearly the entire distribution of the overall score. We 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 is nearly zero at all levels of the scoring threshold (Figure 4 Panel b). Yet 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 supports our study's findings, showing our results to be generalizable: female names associated with music compositions influence reviewers' preferences throughout nearly the entire distribution of 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. Senior faculty 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 test whether there is a statistically significant difference in evaluation scores between younger faculty and elder faculty reviewers; Panel C in Table 5 shows that this difference is only at the marginal line of statistically significance in overall evaluation (P-value<0.1), and not significant in structured evaluation.

Table 6 presents results showing scoring separated out among participating faculty reviewers of differing tenure status: adjunct instructor, assistant professor, associate professor, and full professor. Gender bias was observed among full professor reviewers. They showed preference for female composers in general evaluations and structured evaluations, and compositions associated with female names received higher ratings in all five musical dimensions (although scoring differences for tempo and orchestration were not statistically significant.). All the differences are presented in Table 6 (See also Appendix Figure A-1). Full professors showed the only statistically significant bias in favor of female composers of the reviewer tenure status groups. 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 general evaluation corresponds to 44% of the overall standard deviation. All the lower-ranked professors, including adjunct professors, assistant professors, and associate professors did not exhibit a statistically significant gender bias. 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 marginal line of statistical significance in overall evaluation (P-value<0.1) and not statistically signifcant in structured evaluation. We note that the patterns between the overall score and the average score (based on the structured evaluation) appear slightly different in Tables 5 and 6. We suspect that evaluators have differing opinions on what they consider to be a good composition. We explain more about the general evaluation, which produced an overall score, versus the structured evaluation in the next section.

E. General versus Structured Evaluation

Hypothesis C predicted that discrimination against female composers would be reduced or eliminated through structured evaluation. However, a preference for female composers emerged, regardless of the evaluation method used. Although the differences in magnitudes between the two measurement deigns designs (0.294 points on the 0–10 scale and 0.182 points on the 1–5 scale) are not directly comparable, the results underscore the persistent biases in structured evaluations.

Within the structured evaluation framework, gender bias appeared evident both in the overall evaluation and across all five music categories, with form and tempo achieving statistical significance at the 10% level. Reviewers who favored compositions associated with female names rated these pieces consistently higher, whether assessing specific musical features, detailed components, or the composition as a whole. The use of structured evaluations did not seem to alter the gender biases of faculty reviewers. Although a minor discrepancy exists between the overall scores and the average scores of structured evaluation in the heterogeneity analysis, the current evidence does not suggest that structured evaluations effectively diminish inherent biases.

One key limitation in our study is our adoption of a within-subjects experimental design over a between-subjects approach when comparing overall against structured evaluation. Ideally, one group of respondents could focus solely on the overall evaluation survey, while another group exclusively focus on the structured variant, facilitating a between-subjects test. Several considerations influenced our choice of a within-subjects test. Primarily, given the distinctiveness of our population, a between-subjects test might further constrain the study's statistical power. Additionally, we refrained from confining respondents solely to the structured evaluation, as composition competitions do not typically employ this method. Moreover, directing reviewers to assess only the overall evaluation would be at odds with our stated invitation purpose to explore evaluation criteria.

We have instituted design measures to reduce the likelihood of score harmonization. Once reviewers progress to the structured section, they cannot revert to the previous page to modify their overall scores. Furthermore, the scales for the scores are intentionally differentiated: the overall evaluation is on a 1-10 scale while the structured evaluation uses a 1-5 scale. This distinction makes it less intuitive for reviewers to directly align the two assessments.[8]

6. Robustness Check

We conducted further robustness checks using different model specifications to see if our results are consistent and valid. Table 7 shows the response time. We thought it unlikely that response times would matter much in driving the results as some professional judges may simply have downloaded the scores and performed the evaluation offline making their response time appear very quick. To test the results' robustness, we simply drop those who have unusually short response times (less than 10% of the bottom tail). We find that these results are almost identical to the main results in Table 2. There is also no significant difference in the response times between our randomization group A and group B.

To assure our external validity, we coded the background information of more than one thousand respondents and non-respondents and test their differences. We are constrained by the basic information available on faculty's website. Based on the results in Table 8, we find that our respondents are very similar to non-respondents in terms of the gender proportions, and whether

[8] In future surveys, switching the order of the overall and structured evaluations could help mitigating the anchoring effect, where faculties might attempt to harmonize the two types of scores.

the faculty's institution is public or private. 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 female favoritism.

To address any concern that the results are driven by outliers, we re-estimate our main coefficients after dropping the observations that have overall scores below 1st and above 99th percentile. The results, shown in Table 9 and Table 10, are consistent with our main results.

7. Discussion

To better understand our findings, we delve into the underlying mechanisms that could explain our results, using different economic models of discrimination. Predominately, the literature identifies two primary sources of discrimination: taste-based discrimination and statistical discrimination. More recently, a new paradigm has emerged, discussing a dynamic model of discrimination, rooted in biased belief, which is relevant to our context (Fryer, 2007; Fryer and Jackson, 2008; Schwartzstein 2014; Bohren et al., 2019; Bohren et al., 2022; Bohren et al., forthcoming). Given the complexity inherent in the discrimination literature, we discuss whether and to what extent our results align with the three overarching frameworks that illuminate the sources of discrimination. In our discussion, we integrate the dominant terminologies associated with these three discrimination types, leaning heavily on the languages from Bohren (2022) due to its framework's clear and fitting alignment with our study. However, we are unable to address all the underlying origins of these biases, such as those arising from stereotype, prejudice or cultural norms.[9]

In summary, while our findings diverge from the classical statistical discrimination model (belief-based with correct belief), they more closely reflect the patterns of the “belief-based” model with incorrect belief. In this later model, a subsequently updated belief reverses the initial bias. Although we have limited information to adjudicate the exact mechanism, we provide highly plausible explanations in detail below.

7.1 Consistency with Belief-based Model (with Incorrect Biased Belief)

In the “belief-based” literature with incorrect biased belief, discrimination is dynamically updated with new information after evaluators receive more signals. There is evidence of settings, for instance, where evaluators hold initial discriminatory beliefs against a group (e.g. women) but quickly update them once they see evidence of good performance. Discrimination could be reduced or even reversed after the discriminated group earns a better reputation. This reversal in Bohren (2022)’s setup is driven by the presence of two types of evaluators. One exhibits a belief-based bias favoring men over women, which generates the initial discrimination. The other type perceives the existence of the first type. With the heterogeneity of those two types, an average reversal of the initial bias can be observed.

As suggested by this literature, it is possible that the evaluators have internalized how difficult it is to “make it” in the music world as a female composer. If this were the case, a previous stage of

[9] Naddeo (2022) explored multiple factors in the U.S. jurisdiction, distinguishing between biases originating from systemic forces and those from direct sources. In our study, we lack the systemic indicators to differentiate these sources. In addition, unlike the U.S. jurisdiction context, the historical gender imbalance in the discipline persists.

selection as well as a "belief update" would have already occurred. In the evaluators' updated belief, the compositions associated with female names are worthy of higher ratings as they are authored by women who have succeeded in becoming music composers. Our evidence of heterogeneity in which senior faculty are more favorable toward female composers than junior faculty would also fit well with the "belief update" mechanism.

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

7.2 Alternative Models: Statistical Discrimination (Belief-based with Correct Belief)

The other model commonly used to explain discrimination is the statistical discrimination model with correct belief. Within this framework, biased beliefs stem from commonly held knowledge based on observable characteristics, such as gender. In the dynamic prediction of this model, the initial biases will be eliminated with sequential information, but these biases would never reverse (Bohren 2022). Contrarily, our empirical results show a reversal gender bias, suggesting that reviewers view female composers as more capable based on historical statistical evidence and thus rate them higher than their male counterparts.[10] This explanation is unlikely the case, as we have ample evidence showing the opposite—unlike with certain musical instrument performance, such as piano or flute, the fields of composition and conducting are associated with a leading role in the orchestra and bear relatively negative stereotypes against women. Bennett et al. (2019) conducted

[10] Statistical discrimination often arises from the *correct* perception that one group, in our case, women composers, has a lower, or more precise, or less variant distribution compared to the other group, in this case, male composers.

a survey among female composers and found that most of them reported experiencing gender-related disadvantage in their composition careers. Many respondents stated that they try to strategically conceal their identity in their professional setting, such as composition competitions. In addition, the statistical discrimination model would predict that discrimination would be mor pronounced among less experienced evaluators than among more experienced evaluators. Our results do not support this common prediction. We therefore contend that our findings are not in line with the classical statistical discrimination model's predictions.

7.3 Alternative Models: Taste-based Discrimination (Preference-based Model)

The taste-based discrimination mechanism was derived from Becker's preference-based discrimination model (Becker, 1957; Charles and Guryan, 2008). This model suggests that employers may favor one group (e.g. white employees over black employees) because they incur disutility from interacting with a different group, despite their equivalent abilities. Empirically, in-group or out-group bias can be viewed as an extended version of Becker's model. This extended version of preference-based model may be consistent with the empirical findings in Bar and Zussman (2020). They found that, in the driving tests, male testers generate utility from interacting with students of the opposite gender, resulting in a higher passing rate among female students. In this regard, our results may exhibit a similar taste-based gender bias, in which male reviewers incur utility by having more female composers winning awards, and thus they prefer female composers in the experiment. However, the preference-based discrimination in a dynamic setting would predict that individuals would always stay the same as the initial biases and will never shift this preference after more information provided regarding the ability of the candidates. In our case, as we see the discrepancy between the historical biases and our observed empirical evidence. We are unable to fully support this argument.

8. Conclusion

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

This study adds to the discrimination literature. Contrary to the gender discrimination prediction, we found a bias favoring female names in professional settings. When we tested the in-group bias hypothesis, we found an out-group bias, in which male reviewers gave higher scores toward female composers. These results were consistent across general and structured evaluations.

Some may be concerned that our results are driven by the experiment demand effect. Faculty members, for instance, could have googled the name of the composition and figured out that this was a study about gender, resulting in the Hawthorne effect with male evaluators giving favored evaluations toward female composers. We contend that this effect is unlikely. The study's compositions are all newly commissioned works with no public information available online, so the actual composers' names cannot be revealed. We also double-checked that our fictitious names were not the same as real composers' names. It is quite common that junior composers or composition students do not yet have a prominent online portfolio. One may also be concerned that the two-female-two-male design in our experiment leads to a subtle signaling effect for diversity. We cannot eliminate this factor but note that this design is not uncommon, as many female composers have advanced as finalists of multiple well-known competitions in recent years.

Our study addresses the literature on women's underrepresentation in the academy, which typically focuses on the STEM fields. By examining an extreme case of gender imbalance in a non-STEM field, we fill a critical gap in the literature. While our unique setting may limit generalizability of our findings, we argue that it is advantageous to examine unique settings to explore gender issues, as other studies have done in the cases of Olympic sport evaluators (Sandberg 2017) and competitive chess game players (Backus et al. 2023). Our study examines a non-STEM field with an extreme unbalanced woman representation that may exhibit a larger stereotype against women than other traditional male-dominated fields, such as physics and math. Despite the high degree of male domination in this field and the professionalism of our designed evaluation process, we unexpectedly observed a favoring effect toward women. 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). While our paper alone cannot answer this question, it offers 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 carefully. 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 and mechanisms behind differential treatment. Discrimination can be unintentional, institutional, historical, or come through mundane everyday interactions. Our study, which focuses on the assessment of sample compositions, observes only one arena for a composers' career advancement, and therefore, does not capture the larger institutional or historical dimensions of the problem. We acknowledge that gender bias may exist 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). Another caveat is that our study is not a hiring experiment, and thus the results have limitations in translating into employment in music composition professions. Finally, our study does not capture temporal changes in this field. If the recent call for diversity is an explanation of our results, it may only "benefit" a small number of composers of a generation, and the broader inequality remains.

With these limitations in mind, our findings suggest that women's underrepresentation in professional composition is not necessarily a result of unfair evaluation in composition competitions. At the very least, demand side discrimination should not be viewed as the only cause. The issue might also stem from the supply side, with women choosing not to apply for or pursue such careers. Our results 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; Card et al., forthcoming[11]), 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, and more a result of "pre-college factors and the subsequent likelihood of majoring in these fields." (Ceci et al., 2014). The applicability of our insights to STEM and other fields, however, is yet to be

[11] Card et al. (forthcoming) examine the selection of Fellows of the Econometric Society. They found that there was gender gap in earlier decades, and the effect flips favoring women in the most recent decade.

determined. Further research is valuable in other aspects of this profession, such as composers' employment, pay, and career advancement, to yield 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?" AEA Papers and Proceedings 108 (2018): 186-90.

Backus, Peter, Maria Cubel, Matej Guid, Santiago Sánchez-Pagés, and Enrique López Mañas. "Gender, competition, and performance: Evidence from chess players." *Quantitative Economics* 14, no. 1 (2023): 349-380.

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): 221-242.

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." *Handbook of Economic Field Experiments* 1 (2017): 309-393.

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 YRP, Marina Gertsberg, and Rachel AJ Pownall. "An empirical analysis of price differences for male and female artists in the global art market." *Journal of Cultural Economics* 46, no. 3 (2022): 543-565.

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

Bohren, J. Aislinn, Peter Hull, and Alex Imas. "Systemic discrimination: Theory and measurement" (2022). Unpublished manuscript.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. "Inaccurate statistical discrimination: An identification problem. " *Review of Economics and Statistics*. (forthcoming)

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.

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

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

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. Accessed August 10, 2023.

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.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. "Gender differences in peer recognition by economists." *Econometrica*, (forthcoming).

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.

Charness, Gary, Anya Samek, and Jeroen van de Ven. "What is considered deception in experimental economics?" *Experimental Economics* 25, no. 2 (2022): 385-412.

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.

Cooper, David J. "A note on deception in economic experiments." *Journal of Wine Economics* 9, no. 2 (2014): 111-114.

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.

Dustan, Andrew, Kristine Koutout, and Greg Leo. "Second-order beliefs and gender." *Journal of Economic Behavior & Organization* 200 (2022): 752-781.

Edvenson, Arianna “Gender and music: can we hear a difference between female and male composers and performers?” (2017). Unpublished manuscript.

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. "A Categorical Model of Cognition and Biased Decision-Making." *B.E. Journal of Theoretical Economics* 8 no. 1 (2008): 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.

Kessler, Judd B., Corinne Low, and Xiaoyue Shan. “Lowering the playing field: Discrimination through sequential spillover effects.” (2022). Unpublished manuscript.

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.

Naddeo, J.J. " Race, Criminal History, and Prosecutor Case Selection: Evidence from a Southern U.S. Jurisdiction" (2022). Unpublished manuscript.

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): 101-115.

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): 199-214.

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.

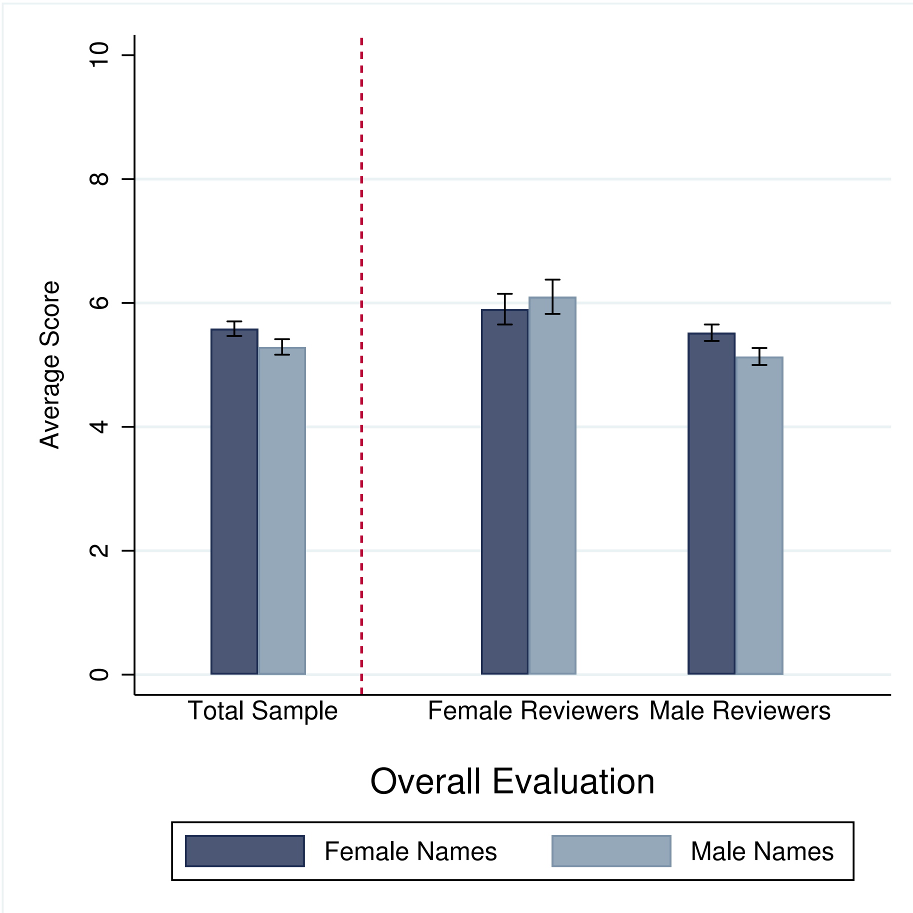
Wu, Alice H. "Gendered language on the economics job market rumors forum." In *AEA Papers and Proceedings*, 108 (2018): 175-79.

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 (2019): 3000-3008.

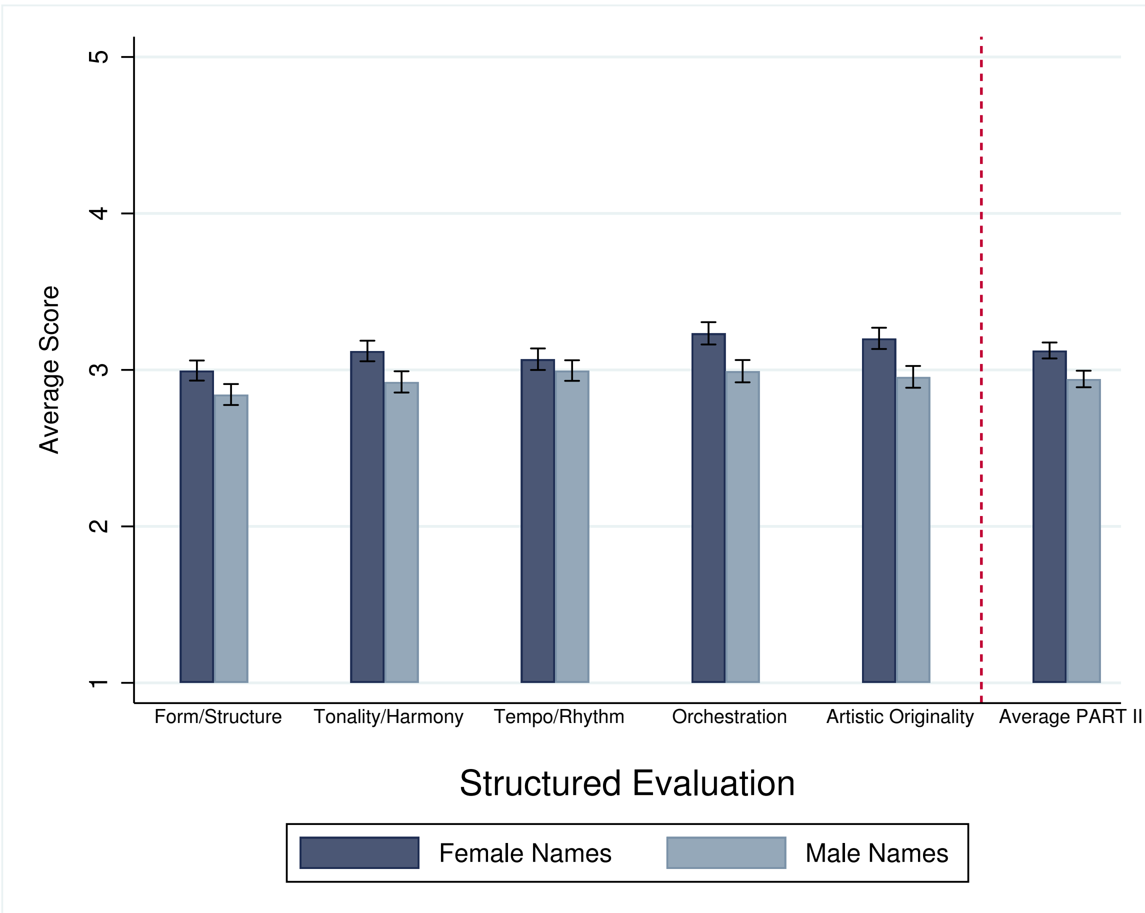
9 Figures and Tables

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.

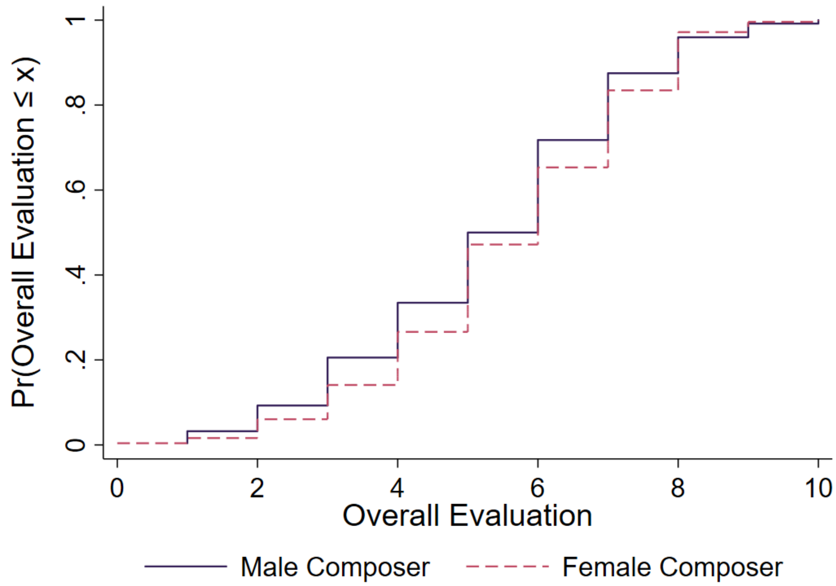
Figure 2: Graphical analysis of structured evaluation



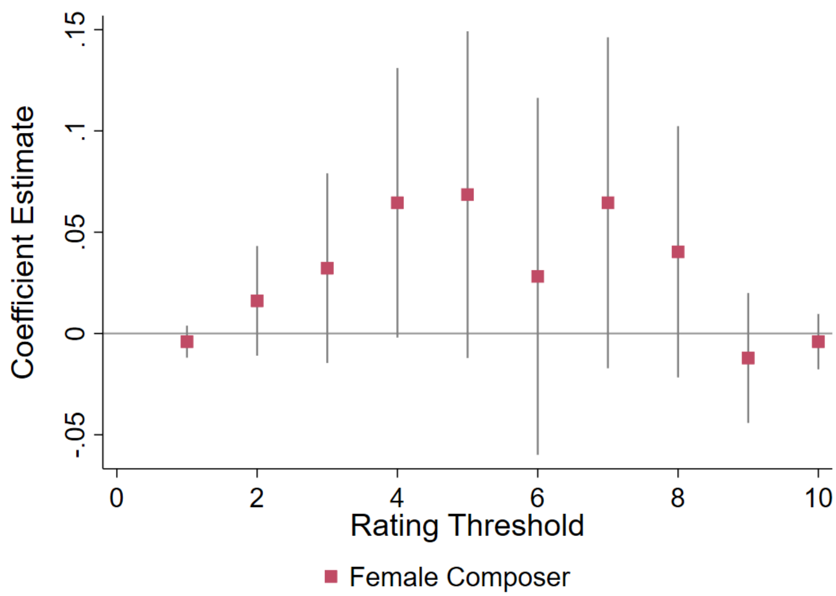
Note: All the bars represent the average scores in Part II structured evaluation (scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong") for compositions associated with female names vs. male names. Compositions are evaluated in five categories.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



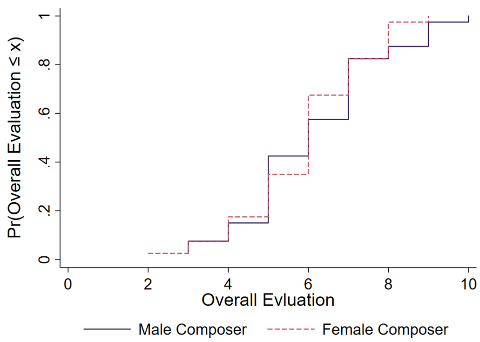
(b) Linear Probability Model for *Female Composers*



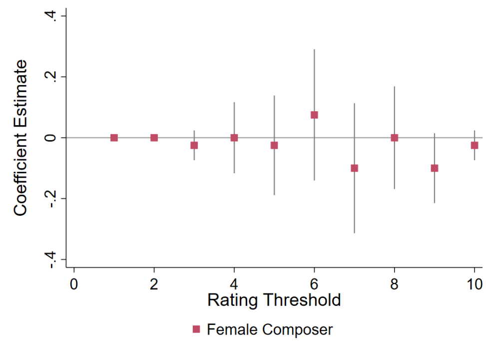
Note: Empirical CDF of *Overall Evaluation Score* is in panel (a). The CDFs show the share of compositions with each characteristic (ex: written by hypothetical male or female composers) with a overall evaluation score less than or equal to each value. The counterfactual callback rates for *female names* 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

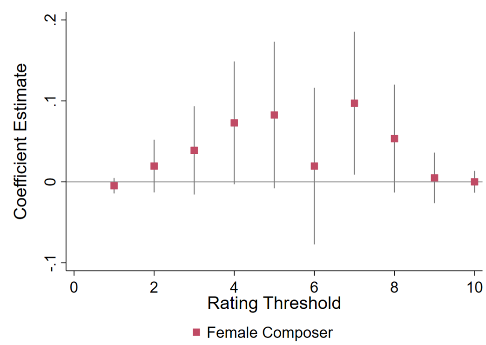


(b) Linear Probability Model for *Female Composers*, Female Reviewers



(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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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.

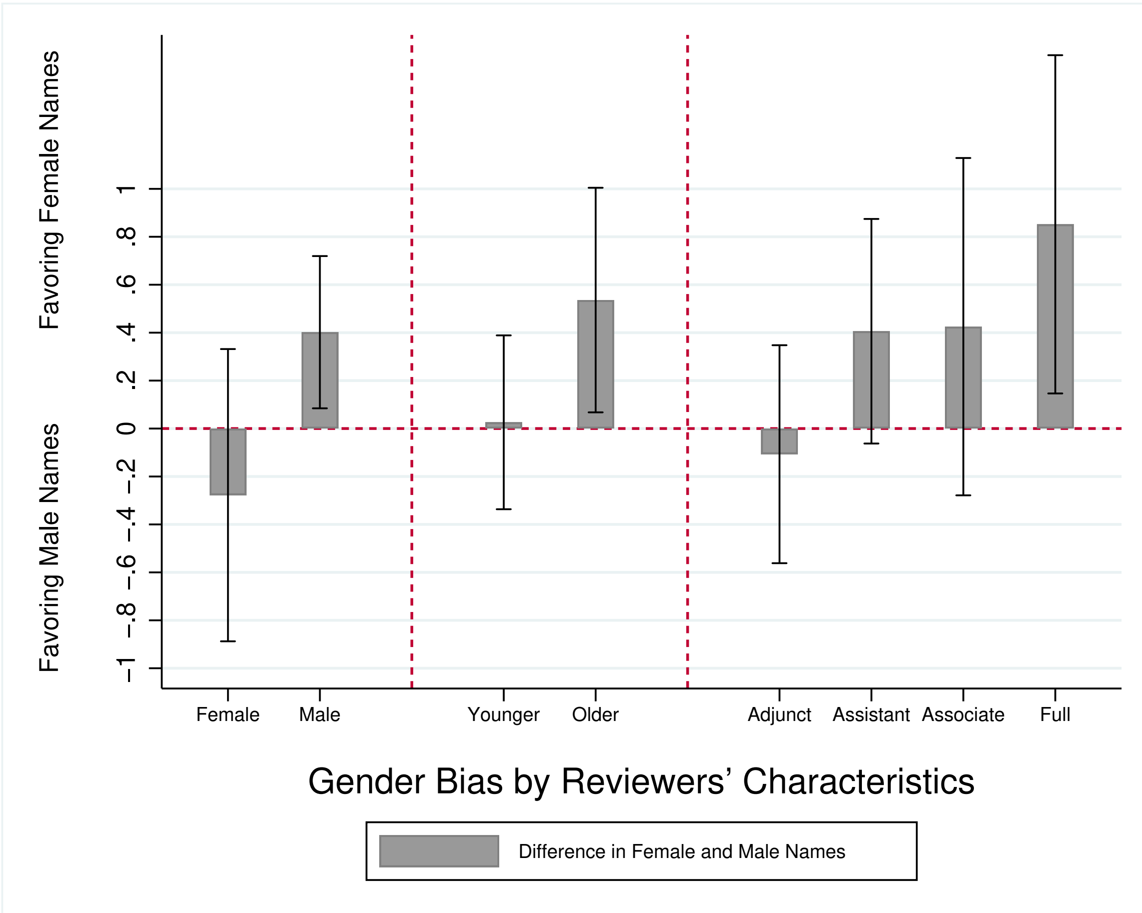
Table 10: Robustness Check: Dropping Outliers (In-group Bias)

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. Part I is the overall evaluation scored on a 0–10 scale. Part II is the structured evaluation scaled from 1 to 5, namely "Extremely Weak", "Somewhat Weak", "Neutral", "Somewhat Strong", "Extremely Strong." 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