ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Gender bias in the evaluation of artists: an empirical analysis of voting behaviour in the Eurovision Song Contest

---

Master Thesis

MSc Economics and Business

Specialisation: Behavioural Economics

Author:

Ludwig Baunach

621956

Supervisor:

Dr. Georg Granic

Second assessor:

Francesco Capozza

July 22, 2022

Abstract

This study analyses voting behaviour in the Eurovision Song Contest from 2016 to 2021. In particular, it investigates how gender of a performing artist affects a country's evaluation of the artist. On average, women got ranked significantly worse than men. The effect is significantly stronger among tele votes compared to jury votes. In contrast to the second hypothesis, gender gaps that capture gender inequalities in countries failed to explain differences in gender effects among votes. Thus, results did not show that countries with wider gender gaps vote differently among genders than countries with smaller gaps. A self-selecting voting base among the ESC viewers could have led to selection bias in the voting data. Moreover, different voting mechanisms for jury and televoters may bias the results. Therefore, it is not possible to conclude that causal effects were observed when measuring gender effects, nor is it possible to conclude that artists are affected by discriminating gender bias in the ESC. Nevertheless, the results support prior research that suggests jury members are less prone to biases and will vote more objectively.

Keywords: Gender Bias, Voting Behaviour, Eurovision Song Contest

Table of Contents

List of Tables

List of Figures

Introduction

In 2021, 183 million viewers followed the Eurovision Song Contest (ESC), making it the biggest live music event worldwide (European Broadcasting Union, 2021). The contest has evolved from a small regional challenge between seven nations in 1956 to an event of significant amplitude. Now, 44 countries compete each year to win the contest. The winner gets elected by vote. Voting rules changed several times, including using jury and public votes. Voting data from each year is published openly on the official contest's website, making it probably the largest non-political election with an open dataset. Due to the rich dataset and its special setup of competing countries, the ESC has been attracting social scientists and economists to study human voting behaviour.

Although the contest's goal is to provide a fair vote for the best musical performance, many assume that other factors impact the final rank of a contestant as well (see Clerides and Stengos, 2012; Dekker, 2007; Fenn et al., 2006; Gatherer, 2006; Ginsburgh and Moreno-Ternero, 2022; Ginsburgh and Noury, 2008; Haan et al., 2005; Spierdijk and Vellekoop, 2009; Stockemer et al., 2018). Understanding voting behaviour allows us to judge whether the goal of electing the best musical performance is achieved. For example, if certain political events that are independent of the artist's performance affect the rankings of an artist, as shown by Spierdijk and Vellekoop (2009), a fair vote for the best musical performance is not achieved. In addition, being aware of voting behaviour can help us discuss fallacies and construct or use fairer voting methods. For example, suppose jury votes are more objective regarding the quality of a song, as mentioned by Haan et al. (2005). In that case, one may argue that artists should only be evaluated by such and not by the public.

Furthermore, analysing ESC voting data can help us understand voting behaviour. It can also provide valuable insights that may be used to understand elections outside the music industry. This matters for a particular reason. In many cases, voting is a form of performance evaluation (e.g. contests) or power- or capital allocation (e.g. politics or funding). As soon as evaluations are not based on quantitative aspects like the shortest time to run 100 meters but on personal judgement like selecting the best job candidate by personal interviews, problems may arise. Suppose people do not evaluate only performance quality but are biased in their votes by other factors such as race, gender, age or cognitive biases in a more general sense. In that case, we

can conclude that the outcomes of these votes are not objective nor rational and are not based on the collective beliefs of quality of performance. On the one side, such findings would raise strong ethical concerns since equal opportunity across those factors would not be achieved. On the other side, economic concerns can be raised. Biases in voting would lead to the consequence that the best performance may not be elected first. Consequently, the allocation of power and social credit would not be objectively optimised in society.

For this reason, analysing voting methods, voting behaviour, and the fallacy of such has been of a long interest in economics. There are many elections and contests to analyse. However, few offer such rich and transparent data over a long period as the ESC. A body of literature has committed to the ESC to test and analyse different aspects of voting. Most of the published papers focus on cultural, geographical and political ties between countries and societies and how they influence voting behaviour (see Clerides and Stengos, 2012; Dekker, 2007; Fenn et al., 2006; Gatherer, 2006; Ginsburgh and Noury, 2008; Spierdijk and Vellekoop, 2009). Others have used the ESC voting data to compare different voting mechanisms and aspects of social choice theory, testing and suggesting different voting mechanisms in the contest (Ginsburgh and Moreno-Ternero, 2022). Some have referred to differences in public versus jury votes (Haan et al., 2005; Clerides and Stengos, 2012). The list goes on, however, misses one major factor, the influence of gender of artists on their ranking. Even though there is a long holding interest in gender bias in many fields, such as medicine (e.g. Hamberg, 2008), clinical psychology (e.g. Garb, 1997), machine learning (e.g. Caliskan et al. 2017), and not to forget performance evaluations (e.g. Goldin and Rouse, 2000; Card et al., 2020), gender bias was not of primary interest in the ESC literature to this point. Using gender as a control variable is common, but none have focused their work on analysing gender effects in the ESC.

This work aims to contribute to the field by making this the main research focus. Gender bias can be hard to grasp. Is there a bias, or do we observe legitimate preferences? This question is challenging and will be discussed along the way and in more detail in the discussion section. Nevertheless, subsequently, gender bias means that the gender of an artist that is evaluated matters independently of their quality of musical performance. For further elaboration, take the research of Goldin and Rouse (2000) as an example. They introduced blind auditions, which consequently increased the probability of female musicians getting ranked higher or hired in symphony orchestras. The experiment showed that, in their case, gender affected the evaluation

independently from the musicians' performances. Similarly, this thesis aims to investigate the effect of an artist's gender on their rankings. This goal leads to the following research question:[1]

How does the gender of a performing artist affect a country's evaluation of the artist in the Eurovision Song Contest in the years 2016-2021?

In this work, I examine the following three points. First, a positive correlational association between the variables Female and Final Rank of an artist is tested.[2] Note that higher ranks indicate worse ranks since rank 1 is best and rank 26 is worst. A positive correlation would suggest that being female is associated with worse (higher) rankings. However, such an observed correlational effect would not provide sufficient evidence to conclude gender bias.

Second, the annually published Global Gender Gap Index (GGGI) is used to test whether there is a positive correlational relationship between the gender of an artist and the degree of gender inequality in the evaluating country captured by the GGGI. Overall, the GGGI is of interest in this research because it provides an explanation for the effect of the gender of an artist on their evaluation in the ESC dataset. Gender bias means that the gender of the person evaluated matters independently of the quality of their performance. Suppose the effect size is independent of the quality of performance and can be partially and significantly explained by the size of a gender gap in a country. In that case, one can conclude that gender inequality in societies affects the evaluation of artists in these societies based on their gender. The interesting and unique part of this analysis is that the evaluated artists are not from the country that evaluates such artists. Thus, societal influences of the voting country do not influence the artists' performance. Therefore, if evaluations differ significantly among the different scales of gender inequalities among voting countries, one can conclude that there is suggestive evidence that inequalities in countries affect the evaluations of artists based on gender, independently of the quality of performance. In that case, suggestive evidence for present gender bias among gender-unequal counties would be provided.

[1] The rules of the contest have changed many times over the years. In 2016, the European Broadcasting Union introduced a new voting mechanism that includes both the jury and the public vote. Therefore, the work will be limited to the years 2016 and after.

[2] Female indicates the gender of an artist. Final Rank captures the combined rank of jury and tele votes.

Last but not least, the effect of gender on the final ranks of artists is compared between jury and public votes. Prior research has suggested that the ESC jury members are less prone to biases than public voters (see Haan et al., 2005; Cleredis and Stengos, 2012). A significant bigger effect of Female among public votes on Final Rank of an artist would indicate that other factors than quality of performance are driving the effects. Such effect would support but not prove a claim of gender bias in voting. In addition, if the contest's goal is to select the best performing artist and jury votes are more objective than tele votes, one may question the use of tele votes in the contest. In 2016, a new voting mechanism that includes both the jury and the public vote was introduced. Analysing the data from 2016 onwards allows comparing jury and tele votes in the respective years. Beforehand, jury and tele votes were compared from different years.

The thesis is structured as follows. First, the ESC and GGGI are formally introduced. Next, the research question, the three main hypothesise, and control variables are derived from the related literature. Further, literature connected to other factors that may influence the results is discussed. The literature review is followed by an introduction of the methodology used to test the research question empirically. Then, results from the empirical analysis are presented and interpreted. Lastly, limitations, implications and external validity of the research are discussed. The final discussion critically reflects on the study and gives directions for further research.

The Eurovision Song Contest

The ESC originated from an Italian music festival that initially intended to test the limits of live television broadcast technology. The contest was held in 1956 for the first time. At this point, seven nations participated. After that, the contest was held annually (except in 2020), and in total, 52 countries participated at least once. The European Broadcasting Union (EBU) organises the contest, and only active members of this union are eligible to participate. However, several states outside the European continent's geographic boundaries are active members of the European Broadcasting Union. Therefore, participation in the contest is not limited to European countries (European Broadcasting Union, n.d. b). Examples of countries outside of Europe are Israel, Cyprus and Morocco.

A maximum of 44 countries are allowed to participate. Each participating country chooses their artists. The contestants can perform in groups (up to 6), duo or solo. Furthermore, each participating artist is free to decide the language in which they will sing. Six countries are automatically pre-qualified for the final of the contest; France, Germany, Italy, Spain and the United Kingdom – and the host country. The host is usually the winner of the previous contest. The remaining countries will compete in two semi-finals. From each semi-final, the best ten performances proceed to the Final. Thus, 26 countries compete in the Final (European Broadcasting Union n.d. a).

In 2016 the voting rules changed. Since then, both professional juries of 5 and televoters (viewers) from each country vote. Beforehand either tele or jury votes were used to determine the contest's winner. The change is particularly interesting because it allows a direct comparison between jury and tele votes. This comparison is incorporated in the third hypothesis of this research. Voting works as follows. After all performances, viewers in each country participating in the contest may vote for their favourite performance. There will be a 15-minute window for televoters to vote. The public can give multiple single votes for multiple countries in that time frame. Note that viewers are not allowed to vote for their own country. Concerning televoting, songs are ranked according to the total number of votes a country receives. Viewers can vote multiple times. On the other side, jury members vote differently. Each jury member ranks each song from 1 to 26. The final jury ranking is determined by the average of the five individual jury member rankings. The tele- and jury rankings are used to calculate the points a country receives. The first place gets 12 points, the second 10, the third 8 and the fourth 7, now decreasing to 0 points with 1-point steps. The remaining will be awarded 0 points.

Consequently, there are two sets of ranks and points from the jury and the public. In the end, the points are taken to nominate the winner. Jury and tele points are equally weighted (European Broadcasting Union, 2022a). Note that points are only given to the first ten ranks. The rest will get 0 points. Thus, the set of ranks contains more accurate information about the overall performance preference. In the end, there are two sets of ranks. One is given by the jury henceforth called Jury Rank, and one is based on tele votes henceforth called Tele Rank. The combined rank of Jury Rank and Tele Rank is called Final Rank.

The Global Gender Gap Report

The GGGI is an index to measure gender equality in countries. This thesis uses the index to test whether societal gender inequality may be associated with gender bias in voting. It has been published annually by the World Economic Forum since 2006. All subsequent explanations in this thesis are based on the methodology section of the Global Gender Gap Report 2021 written by Crotti et al. (2021, p. 73 ff.). There are three underlying concepts to the index. First, it aims to measure gender gaps in countries' access to resources and opportunities rather than the actual levels of resources and opportunities. Second, it evaluates countries based on outcomes rather than inputs or means. It aims to give a snapshot of where men and women stand concerning fundamental outcome indicators related to basic rights, including economic participation, education, health and political empowerment. The third distinguishing feature of the GGGI is that it ranks according to quality instead of women's empowerment. That means that it does not reward (nor punish) inequality in outcome in favour of women. For example, a country that has higher enrolments for girls than for boys in middle school will score equal to a country where enrolments are even. The World Economic Forum uses four sub-indices to calculate the overall GGGI.

First is economic participation and opportunity. This subindex contains three concepts: the participation gap, the remuneration gap and the advancement gap. First, the participation gap is captured using differences in labour force participation rates. Second, the remuneration gap is measured by differences in earnings. Finally, the advancement gap is captured through the ratio of women and men in higher labour positions.

The second is educational attainment. It captures the gap in access to education through the ratios of women and men in primary-, secondary- and tertiary-level education. A longer-term view of the country's ability to educate women and men in equal numbers is captured through ratios of literacy rates.

Third, health and survival. This subindex captures differences between women's and men's health using two indicators. The first is the sex ratio at birth. It relates to the phenomenon of "missing women", prevalent in many countries with strong son preference. The second ratio is the differences in life expectancy.

Last but not least is political empowerment. This subindex captures the gender gap at the highest level of political decision-making. Ratios of women and men in ministerial and parliamentary positions and ratios of women and men in terms of years in executive women are considered.

All subindices are converted into ratios that work as a foundation to calculate the subindex scores. The overall score is calculated by taking a weighted average across the subindices. For all subindexes and the overall score, the highest possible score is 1 (gender parity), and the lowest possible score is 0 (imparity). For example, in 2021, the GGGI score of the Czech Republic was 0.71, and Norway had a score of 0.85. Hence, Norway closed the gender gap by 85%, whereas the Czech Republic closed the gender gap by only 71%. The differences indicate that Norway is more gender equal than the Czech Republic.

Overall, the GGGI is of interest in this research because it provides a possible explanation for the effect of the variable gender in the ESC dataset. Gender bias means that the gender of the person evaluated matters independently of the quality of their performance. Suppose the effect size is independent of the quality of performance and can be partially and significantly explained by the size of a gender gap in a country. In that case, one can conclude that gender inequality in societies affects the evaluation of artists in these societies based on their gender. The interesting and unique part of this analysis is that the evaluated artists are not from the country that evaluates such artists. Thus, societal influences of the voting country do not influence the artists' performance. Therefore, if evaluations differ significantly among the different scales of gender inequalities among voting countries, one can conclude that there is suggestive evidence that inequalities in countries affect the evaluations of artists based on gender, independently of the quality of performance.

Literature Review

The subsequent chapter will discuss literature relevant to the research. In general, the structure of the literature review is based on the overall structure of the research. Therefore, it starts with the literature closely related to the main research question and then derives the main hypotheses. Furthermore, other biases that have been observed in voting are connected to the ESC and are relevant for the statistical analysis are discussed. In this regard, control variables are explained, and possible limitations are introduced.

There is a long holding interest in gender bias in many fields, such as medicine (e.g. Hamberg, 2008), clinical psychology (e.g. Garb, 1997), machine learning (e.g. Caliskan et al., 2017), and not to forget performance evaluations (e.g. Card et al., 2020). However, gender bias was not of primary interest in the ESC literature to this point. This is the case even though there are observed gender biases in the music industry regarding the performance evaluation of musicians. For example, in a labour economic experiment, Goldin and Rouse (2000) found evidence that introducing blind auditions increases the probability of female musicians getting ranked higher or hired in symphony orchestras. The experiment showed that, in their case, gender affected the evaluation independently from the musicians' performances. Furthermore, in a psychological study, Colley and North (2003) link male-dominated pop music to an anti-female bias-based stereotyping effect. As mentioned before, gender bias means that the gender of an artist that is evaluated matters independently of their quality of musical performance, as shown in the experiment of Golding and Rouse (2000). Regarding the ESC, there are indications in the literature that gender may affect the final ranks of artists. Clerides and Stengos (2012) analyse affinity factors in the votes of the ESC. Besides their main results, they highlight the importance of other factors that influence voting patterns, such as the gender of a performing artist. However, at the given time, no research explicitly aims to investigate the effect of gender on the ESC. This research takes this finding as motivation to test whether there are associations between the variable gender and the final rank of an artist, leading to the following research question.[3]

[3] The rules of the contest have changed many times over the years. In 2016, the European Broadcasting Union introduced a new voting mechanism that includes both the jury and the public vote. Therefore, the work will be limited to the years 2016 and after.

How does the gender of a performing artist affect a country's evaluation of the artist in the Eurovision Song Contest in the years 2016-2021?

To address the research question, in total, three hypotheses are evaluated. Based on the discussed literature regarding gender bias in the music industry and gender effects in the ESC, the first hypothesis is formulated. Since gender differences usually favour men over women, it is expected to observe similar effects in the ESC data. This leads to the first hypothesis in this research.

Hypothesis 1: On average, female artists receive worse rankings in the combined vote of the jury and public.

As mentioned before, gender bias is of interest in many fields. In some cases, the GGGI is used to test gender differences, understand gender relations or discuss gender stereotypes, supporting the relevance and possible applications of the index. As a reminder, the GGGI captures gender inequalities in societies and uses sub-indices to measure gender differences among the four categories economics, education, health, and politics. In psychology, for example, the book of Rudman and Glick (2021) uses the GGGI to discuss the social psychology of gender in a more conceptional form. Likewise, Koenig et al. (2011) use the GGGI and their results to understand better and interpret their findings regarding masculine leader stereotypes. More importantly, the GGGI is also used in more empirical analysis in economic and educational research.[4] A well-known paper by Guiso et al. (2008) analysed gender differences in math and reading tests with indicators of gender equality (one being the GGGI). One result shows that when comparing Turkey (GGGI = 0.59) and Sweden (GGGI = 0.81), they observe an increase in the mean score performance of girls relative to boys in reading by 18 points, which almost doubles Turkey's reading gap in favour of girls. Their results suggest that the gender gap in math, although it historically favours boys, disappears in more gender-equal societies. Freyer and Levitt (2010) replicated Guiso et al. (2008) results when analysing Pisa data. They conclude that there is a strong positive association between the GGGI measure of female opportunity and the relative performance of girls in math. The methods of Freyer and Levit (2010) and Guiso et al. (2008) are taken as inspiration to use the GGGI to test gender differences in

[4] As a reminder, a GGGI value of 1 indicates gender parity, and the lowest possible score of 0 indicates imparity between males and females.

the ESC in this work. There may be many possible reasons why gender may have an association with the rank that a country gives to an artist. However, suppose the GGGI can predict a part of the effect of gender on the Final Rank. In that case, there is correlational evidence that some portion of the effect of gender is associated with gender inequality in the voting country. This is the case if countries with a wide gender gap evaluate females on average worse than countries with a tight gap and vice versa. Hence, this thesis investigates the possibility that societal gender inequality may associate with gender bias in voting. Again, based on prior findings discussed above, it is reasonable to assume that a wider gap would benefit males. The second hypothesis is derived from this thought and is the following.

Hypothesis 2: On average, females get ranked worse with decreasing GGGI in both the combined vote of the jury and the public.

To further investigate gender bias in the ESC, one can analyse differences between jury and tele votes. Haan et al. (2005) made an interesting finding in the data from the ESC. They show that experts are better judges of quality in the sense that the outcome of finales judged by experts is less sensitive to factors unrelated to quality than the outcome of finals judged by public opinion. Clerides and Stengos (2012) confirm these results in their findings. Furthermore, one can assume that experts are closer to an objective quality ranking than tele voters. According to these findings, a gender bias should be less prominent in the jury vote than in the gender vote. This realisation leads to the third and final hypothesis.

Hypothesis 3: The effect of being female on rank is significantly bigger for Tele Rank than Jury Rank.

This research uses ESC data starting from 2016, which creates a unique opportunity to compare jury versus tele votes for each country. The rules of the contest changed in 2016. Since then, every country provides consistently three sets of ranks, Jury, Tele and Final Rank. This differs from Haan et al. (2005) and Clerides and Stengos (2012), who had to pool votes from different years with different voting methods to test differences.[5]

[5] Voting methods were inconsistent because televoting was introduced as an option in 1998 and was soon adopted by all participating countries.

To this day, many other influences and biases than gender bias have been discussed in social choice and ESC literature. Some of them may affect the results in the final analysis or at least raise concerns that results are biased in any direction and have to be discussed accordingly. One of these biases is the order effect. Fortunately, the order of artists in the contest is tracked and published. Order effects and their influences on elections and votes have been discussed for a long time. For example, Robson and Walsh (1974) find evidence that the order of candidates in the 1973 Irish election significantly affected the candidate's votes. In the Queen Elisabeth Piano Contest, which takes place in Brussels every four years, significant order effects have been observed (for instance, Flores and Ginsburgh, 1996; Glejser and Heyndels, 2001). Although the order of pianists was chosen randomly, pianists that performed last were ranked significantly higher. Cleredis and Stengos (2012) found similar order effects in the ESC voting data. Again, songs that performed later (closer to the voting window) performed better than earlier songs. In this context, Harris et al. (2022) explains position effects in individual choice with the behavioural concept of choice fatigue. Based on the given research, the control variable Order has been included in the regression analysis.

Another effect that can be controlled is the language of a song. Clerides and Stengos (2012) observe significant effects in their control variable language of a song. On this matter, Spierdijk and Vellekoop (2009) conclude that, on average, countries prefer songs in a related language and Blangiardo and Baio (2014) report results that suggest that artists singing in a language different from English are generally scored lower than those singing in English.[6] Because of these findings, the control variable English is used in this research's regression analysis to control for songs that are not sung in English.

Clerides and Stengos (2012) not only mentioned the effects of gender, jury votes and language but also found suggestive evidence that being the host of a contest may positively affect the rank of an artist. Host countries can be easily identified, and the effect is controlled for that reason.

[6] On this basis, the variable English is included as one of the control variables later in the regression analysis.

Control variables in the research of Blangiardo and Baio (2014) indicate that female solo artists get higher scores than group performances. However, the effect is not observed for male solo artists. Thus, the effect of the number of performing artists on stage may depend on gender. For this reason, the control variable Solo is used in the regression analysis. It indicates solo, duo and group performances.

Other effects that have not been controlled for in this research but may influence the results were also discussed in prior literature. For example, as mentioned above, the order of oneself matters and the order of others relative to oneself. In other words, contrast effects may occur during contests. A contrast effect is the improvement or diminishment of perception, cognition or performance due to exposure to a stimulus of greater or lesser value. There is a large body of literature on contrast effects, for example, in psychology, law or economics (Herr et al., 1983; Lynch et al., 1991; Tversky and Simonson, 1993; Kelman et al., 1996). Ginsburgh and Moreno-Ternero (2022) show in a cross-sectional analysis of the 2021 ESC that being surrounded by bad performers may enhance one's performance or the perception of those who have to judge the performance.

Furthermore, Ginsburgh and Moreno-Ternero (2022) comment on the short time frame between the last performance in the ESC and the end of the voting window. In total, tele voters have a time frame of 15 minutes to make the final decision. They raise the concern that speed-accuracy trade-offs could affect the quality of judgement or decision-making. Economic research has discussed this matter with contradicting results. For instance, Kocher and Sutter (2006) have found no loss of quality in decision-making under time pressure in a beauty game. On the other side, Fehr and Rangel (2011) observed more noisy decision-making under time pressure, questioning such quality. Nonetheless, Ginsburgh's and Moreno-Ternero's concerns seem to be valid in the setting of the ESC.

Another example of influences that have not been controlled for is addressed by Sigelman and Sigelman (1982). They used an experimental approach in a simulated mayoral election to test sexism, racism and ageism in elections. They found strong effects for age and similarity effects between voters and contestants. Although age may also affect outcomes in the ESC, it has not been addressed this far.

Effects of similarities on the other side have been discussed. Regarding contests, Coupe et al. (2018) found suggestive evidence for similarity biases in performance evaluations by analysing votes for the FIFA Ballon d'Or award. The results suggest closer ties between jury members and candidates lead to more apparent biases. Their basic specification suggests that a candidate affiliated with the same national team as the jury member will be 10-20 percentage points more likely to be chosen best player. This phenomenon is related to in-group bias, which is the tendency for us to prefer members of our group while opposing members of outside groups. This effect may influence elections such as the 2008 U.S. presidential election (Rand et al., 2009). The ESC offers the possibility to test for similar effects on national levels. Most of the research regarding ESC voting behaviour has analysed the effects of voting blocs, cultural similarities, affinity and political voting. For example, Dekker (2007) determines five blocks of friendship networks that exchange votes (Eastern, Nordic, Balkan, Eastern Mediterranean, and Western). However, immigrant groups influence the results by voting for their home county. Gatherer (2006) goes even further by concluding that voting blocs in the 1990s have crucially affected the final results on at least two occasions. In contrast to prior literature, Spierdijk and Vellekoop (2009) cannot find any evidence for regional bloc voting, although they uncover significant geographical patterns. Instead, they suggest political voting. Ginsburgh and Noury (2008), on the other side, cannot find evidence that voting behaviour mimics political conflicts and friendships. Fenn et al. (2006) find vigorous vote exchanges among some neighbours. For example, between Cyprus and Greece in the years 1992-2003. Although many neighbours have strong cultural ties, there are exceptions to the rule, such as relations between Spain and France or even counter-examples of rivalries among neighbours. Further, it is hard to state biases when finding similarities in voting among countries or voting blocs. Similar cultural influences might cause similar tastes in music. Thus, neighbouring states can have cultural ties or rivalry affecting voting behaviour and should be considered for that reason. However, this proxy should be taken with caution to predict outcomes. Nonetheless, Clerides and Stengos (2012) find that awarding points to songs goes beyond rewarding the quality of a song. They have identified affinity factors that significantly influence the results through voting clusters. They state that cultural, geographic, economic and political factors are significant determinants of point exchanges.

In this research, however, cultural and geographic factors can be accounted for in a fixed effect approach. If cultural and geographic factors are time-invariant between voting pairs, they can be eliminated using year-fixed effects. However, political and economic voting is probably not a time-invariant component and may be unique for each unit-time observation. In other words, political and economic voting may differ each year. In a certain year, political or economic events may influence countries' ties positively; in another year, that same country pair may have disagreements. Based on the discussed research regarding economic and political voting, one should remember that such influences may affect the results.

Method

The subsequent chapter describes data sources and cleaning. Further, descriptive statistics are provided. Next, the ordered logit model is formally introduced. Last but not least, the methodology of testing the three hypotheses is explained.

Data

The data was manually collected from two sources. First, all ESC-related data is readably available on the official European Broadcasting Union's website (European Broadcasting Union, 2022b). The data includes detailed information about voting, participating contestants and their songs. Second, all data related to the GGGI is publicly available in the annual Global Gender Gap Report on the World Economic Forum's website (World Economic Forum, 2022). The data was gathered from the named websites and later joined together.

Regarding the ESC, data between 2016 and 2021 was considered. As mentioned before, using the data from 2016 and onwards allows for comparing jury and tele vote for the same years. Regarding the GGGI, the years 2015-2021 were considered. The GGGI from 2015 was used to match the ESC in the year 2016. The index scores are always published at the end of a corresponding year. Therefore, the previous year's score is closest to the contest date.[7] Thus,

[7] Note that the titles of the Global Gender Gap Report are misleading. For example, in the years 2015-2018, the year in the report's title corresponds to the end of a publishing year. In contrast, the report published at the end of 2019 is given a title with the year 2020. The 2021 report, on the other side, was published in March 2021. Therefore, it can be used to explain the contest in 2021, which was held end of May 2021.

the year of GGGI was matched with the corresponding year a contest was held. In other words, the GGGI is lagged by one year relative to the year of the ESC contest.

In the years 2016 to 2021, 26 countries performed each final. However, in 11 cases, the performance of a duo or a group was mixed in gender. Due to consistency, the observations with mixed-gender performances were deleted from the dataset.[8] In addition, some GGGI scores were incoherent or unavailable (North Macedonia and San Marino), resulting in missing values and being deleted from the dataset. Consequently, the original 5227 observations were reduced to 4547. This dataset was used to test hypotheses one and two. In addition, indicator variables were added to indicate jury or tele vote to test hypothesis three (differences between jury and tele vote). As a result, the observation count was doubled to 9094 observations because now, every year and country pair has two ranks, one for the jury and one for the tele vote, instead of one aggregated rank.

Descriptive Statistics

The data is an unbalanced panel of rankings by each country of all other countries participating in the final. A unit is uniquely identified by the triple: giver-receiver-year. Note that each pair can appear twice a year. For example, the voting pair (A, B) appears once when A ranks B and once when B ranks A. However, the order of a particular pair does not matter since the gender of the ranked artist matters, not the relationship between the countries. For example, Germany will rank 25 other countries. In this regard, Germany-France and Germany-Albania will be considered two exclusive observations. It is assumed that countries' affinity for each other will be constant over the years and will be part of the unit heterogeneity. Under this assumption, using year-fixed effects in the statistical model will allow the cancellation of the effect of affinity between countries. Table 1 shows descriptive statistics of the variables used in the regression analysis. The main explanatory variable is $\text{Female}_{i,t}$. It is a binary variable holding 1 if the gender of an artist performing for country *i* at the ESC contests in year *t* is female and 0

[8] There are two reasons why mixed-gender performances were deleted from the dataset. First, there are different distributions among different groups based on gender. Groups can have up to 6 members. Due to the smaller amount of group observations, it would be challenging to control for different group-gender constellations. Second, it is hard to tell which gender was most influential during the performance. For example, questions about the screen time for each gender would question the result. Therefore, mixed-gender performances were deleted to analyse the isolated effect of gender on performance evaluations.

if it is male. The gender of an artist was manually assigned according to the pronouns used on the official ESC website. $GGGI_{i,t-1}$ is the global gender gap score (ranges between 0-1) of a country that votes for a country *i*. Figure 1 shows the distribution of the GGGI among the countries that voted in the years 2016 up until 2021. In total, 253 unique country-year observations had no missing GGGI scores. The distribution is skewed to the right. The skewness of the GGGI may be a problem in regression analysis because we need enough observations in the extreme to test whether low GGGI countries evaluate females differently compared to high GGGI countries. A skewed distribution is asymmetric and thus different from a normal distribution. In the case of a positive or right skew, the right tail is longer. Therefore, the mass of the distribution is concentrated on the left of the figure. When observing the x-axis of Figure 1, one can see that more observations are clustered at the lower end of the GGGI scores in this sample (0.65) than at the upper end (0.9). It would be better to have a more normal or uniform distribution to find significant differences between lower and higher values. The number of observations impacts the size of confidence intervals. Since there are more observations at the lower end than the upper end of the distribution, it will be harder to obtain significant differences between lower and higher values.

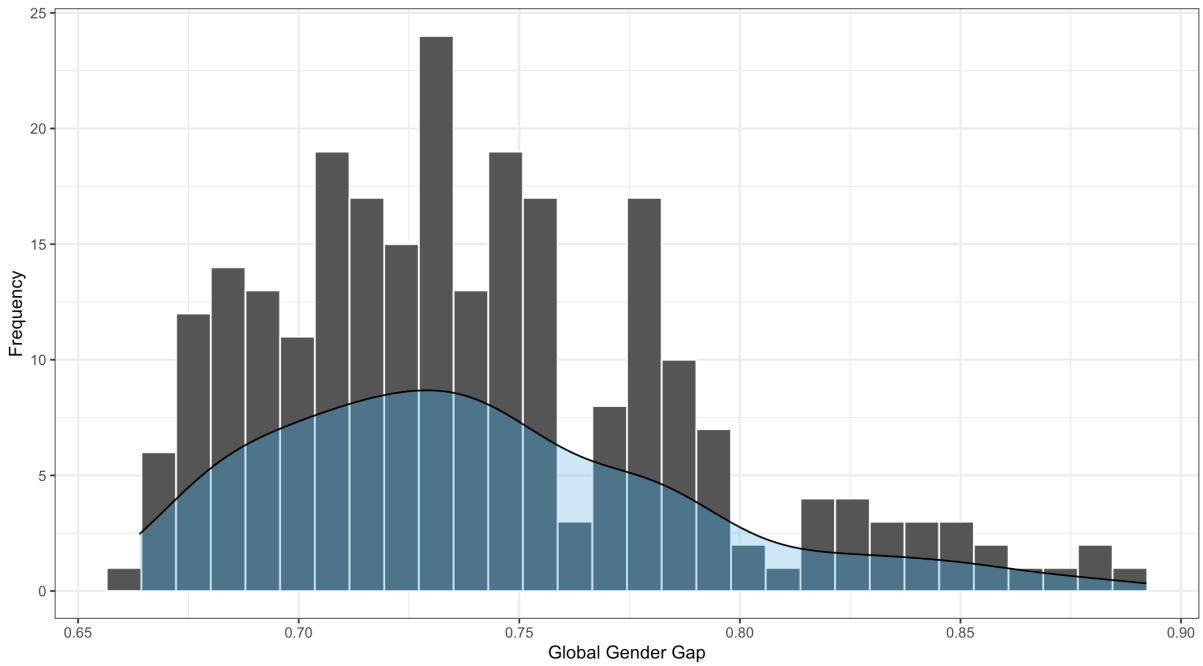

Figure 1: Distribution of GGGI

Note: The graph displays the distribution of GGGI scores of countries that have voted in the ESC between 2016 and 2021 (N = 253).

In addition to the explanatory variables, several observable characteristics were collected. $\text{English}_{i,t}$, $\text{Order}_{i,t}$, $\text{Solo}_{i,t}$ and $\text{Host}_{i,t}$. All characteristics are expected to affect the $\text{FinalRank}_{i,t}$ of an artist of country *i* in year *t* based on findings in prior research and were discussed in the literature review. First, $\text{English}_{i,t}$ is a binary variable with the value 1 if the language of a song performed by country *i* is English and 0 if it is performed in another language in year *t*. Next, $\text{Order}_{i,t}$ indicates the order in which the countries performed in a specific year. For example, if country *i* performed third in year *t*, the value would be 3. Then, $\text{Solo}_{i,t}$ is a categorical variable indicating if country *i* performed solo (1), in duos (2) or as a group (3) in year *t*. Finally, the variable $\text{Host}_{i,t}$ is a binary variable with the value 1 if country *i* is the host country in year *t* and 0 if it is not.[9] Intuitively, the mean rank and order of the variables Final Rank and Order should be between 13 and 13.5 because countries that participate in the finals cannot vote for themselves and will provide 25 ranks and countries that did not participate in the final provide 26 ranks. The average of 25 ranks is 13, and the average of 26 is 13.5.[10] As explained earlier Final Rank is the combined rank of Jury and Tele Rank. However, as mentioned before, some observations had to be dropped due to missing values or mixed-gender performances. For example, the average Final Rank of 13.6 in Table 1 indicates that lower ranks and lower orders were dropped. However, the missing observations should have no association with the effect of Female.

Table 1: Descriptive Statistics

Variable	N	Mean	Std. dev.	Min.	Max.
Final Rank	4547	13.60	7.57	1	26
Female	4547	0.48	0.50	0	1
English	4547	0.76	0.43	0	1
Order	4547	13.54	7.51	1	26
Solo	4547	1.24	0.62	1	3
Host	4547	0.04	0.20	0	1
GGGI	4547	0.74	0.05	0.66	0.89

[9] In the following, variables will be named by name without indications of country and year. The variables $\text{Female}_{i,t}$, $\text{GGGI}_{i,t-1}$, $\text{English}_{i,t}$, $\text{Order}_{i,t}$, $\text{Solo}_{i,t}$, $\text{Host}_{i,t}$ and $\text{FinalRank}_{i,t}$ become Female, GGGI, English, Order, Solo, Host and Final Rank.

[10] The average of 25 ranks is calculated as follows: (1+2+…+24+25) / 25 = 13. The average of 26 ranks is calculated as follows: (1+2+…+25+26)/26 =13.5

Next, measures of association between Female and control variables (GGGI, English, Order, Solo and Host) are reported. Testing for association between the variables is useful for two reasons. On the one hand, it is possible to assess whether the variables provide independent information. On the other hand, it is possible to test the absence of multicollinearity, one of the assumptions in an ordered logit model. As most variables are categorical, standard measures such as Pearson's correlation are inappropriate. Instead, Cramér's V is reported. It provides a measure of association between 0 (no association) and 1 (perfect association). The two continuous variables, Order and GGGI, are transformed into categories by forming quantiles to calculate Cramér's V. Table 2 reports the results.

The Cramér's V between the variables ranges from 0.0021 to 0.2477. When using Pearson's correlation coefficient, the general rule of thumb is that if the correlation coefficient between two variables is greater than 0.8 or 0.9, multicollinearity becomes a problem (Senaviratna and Cooray, 2019). Thus, Cramér's V values below 0.25 do not raise concerns about multicollinearity. Therefore, the assumption of no multicollinearity is not violated.

Table 2: Correlation matrix of independent variables

Variable	Female	GGGI_Q	English	Order_Q	Solo	Host
Female	1					
GGGI_Q	0.0157	1				
English	0.0805	0.0138	1			
Order_Q	0.1620	0.0030	0.2297	1		
Solo	0.2477	0.0170	0.0172	0.1116	1	
Host	0.1129	0.0021	0.0167	0.1489	0.0824	1

Note: This Table provides a measure of association, Cramér's V, between gender, GGGI and the control variables. The range of association bounds between 0 (no association) and 1 (perfect association). Continuous variables have been transformed into categories by forming quantiles (indicated by "_Q").

Regression Model

The main focus of this research is the variable Female and how it affects the Finale Rank of an artist. As mentioned above, three hypothesises are formulated in this regard. Each comes from a different angle. Each hypothesis can be later evaluated along with the economic framework. Although different aspects are addressed in the hypotheses, all three are tested with the same statistical methods. The main model is an ordered logit model. The ordered logit model is more appropriate than linear models because the variable Final Rank is on an ordinal scale. Foundational research on regression models for ordinal data includes the work of McKelvey and Zavoina (1975) and later McCullagh (1980).

Following McCullagh (1980), the ordered logit model can be defined as follows. Let Y be the response in the range 1, ..., k, with $k \geq 2$ and let $\gamma_j = pr(Y \leq j \mid x)$ be the cumulative response probability depending on the covariates x. τ_j are the cut points in the distribution. The model can be derived from a linear logistic model.

$$logit(\gamma_j) = logit\left[\frac{\gamma_j}{1 - \gamma_j}\right] = \tau_j - \beta^T X$$

Long and Freese (2006) describe the measures in a simpler form. In general, the model predicts the following:

$$\Pr(Rank = m \mid X_i) = F(\tau_m - X\beta) - F(\tau_{m-1} - X\beta)$$

For example, $\gamma_j = m$ and X_i being the vector of independent variables, the ordinal logit model calculates the probability of being in a certain rank m depending on the vector of independent variables. In the case of this research, odd ratios are used to interpret the coefficients. They are obtained by exponentiating the ordered logit coefficients. Odds ratios are interpreted in the following way. Ratios of 1 or close to 1 indicate that the odds of a rank among females are the same or similar to the odds of a rank among males. Greater than 1 indicates that the odds of being in a specific rank are greater for females than males and vice versa. Note that odd ratios cannot be interpreted as risk ratios.

In addition to the ordered logit model, two linear models are run: a pooled OLS regression and a fixed effects model. In this case, the dependent variable rank is assumed to be continuous. Linear models are used as robustness tests. Furthermore, all three models are used to test the three main hypothesise. The following chapter will introduce the corresponding statistical hypothesis.

Hypothesis Testing

All three regression models (ordered logit, OLS and fixed effects) take Final Rank as a dependent variable and use Female as the main explanatory variable. Further, the control variables English, Order, Solo and Host are added.

Hypothesis 1 says the following. On average, female artists receive worse rankings in the combined vote of the jury and public. This translates into the following statistical null hypothesis.

H0: The effect of Female on Final Rank is 0.
H1: The effect of Female on Final Rank is greater than 0.

Worse rankings correspond to a higher value in Final Rank since rank 1 is best and rank 26 is worst. Thus, results showing significant positive effects of Female on Final Rank support the first hypothesis.

Hypothesis 2 states the following. On average, females get ranked worse with decreasing GGGI in both the combined vote of the jury and the public. This translates into the following statistical null hypothesis.

*H0: The effect of Female*GGGI on Final Rank is 0*
*H1: The effect of Female*GGGI on Final Rank is smaller than 0*

Female take the value 1 if the gender of an artist is female. The GGGI indicates diminishing gender inequalities with higher values. To support the second hypothesis, the interaction term decreases for females and higher GGGI values. That means that Final Rank of females decrease (improve) with a smaller gender gap.

Finally, Hypothesis 3 says the following. The effect of gender on rank is significantly bigger for tele rank compared to jury rank. JuryRank indicates whether a certain Rank originates from jury (JuryRank = 1) or tele (JuryRank = 0) vote. Therefore, creating an interaction term between JuryRank can indicate the difference between the two groups. This translates into the following statistical null hypothesis.

*H0: The effect of Female*JuryRank on Rank is equal to 0*
*H1: The effect of Female*JuryRank on Rank is equal lower than 0*

Results

The subsequent chapter will present the regression results along the three introduced hypotheses. In addition, visualisations are displayed to improve interpretations of the results.

Hypothesis 1

First, Figure 2 shows the average Final Rank of female and male artists in the ESC from 2016 to 2021. 2020 is missing as the event was cancelled due to the COVID-19 pandemic. The time series shows much higher average ranks for females in the years 2016 to 2019. Over the years, the difference changes between approximately 1 and 4.2 ranks. Rank 1 is the best, and rank 26 is the worst. Thus, the time series is in line with the first hypothesis, that females are ranked on average worse than men in the ESC. Though, in 2021 males were ranked worse than females in the ESC.

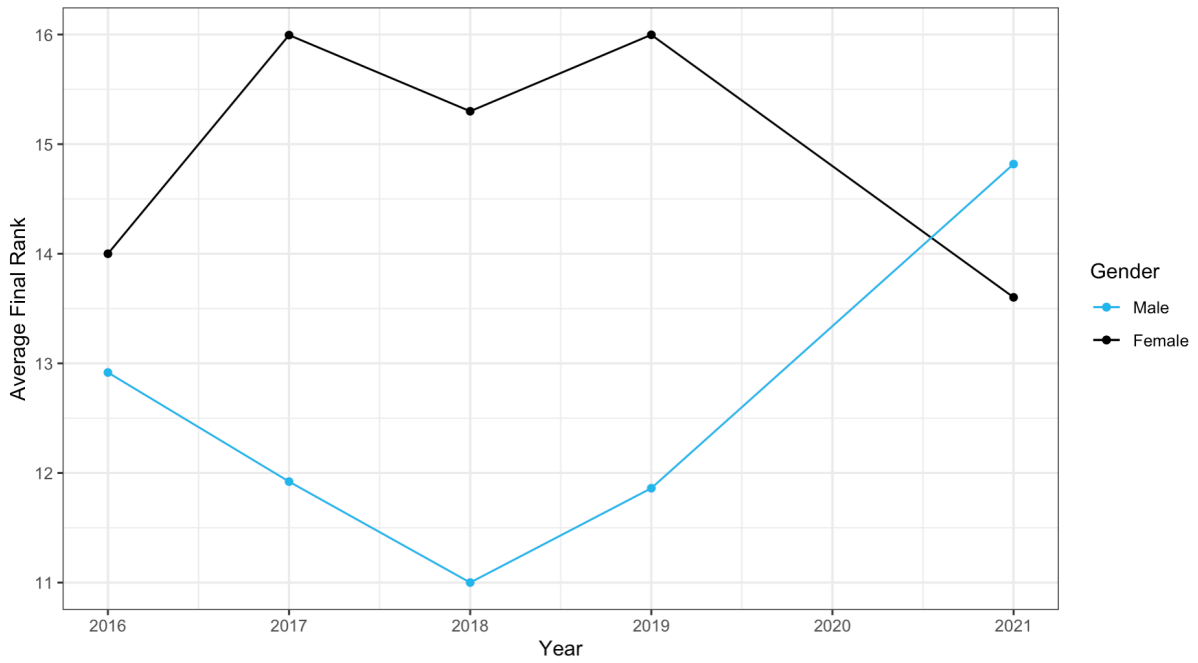

Figure 2: Time Series of Average Final Rank between Male and Female.

Note: 2020 is missing as the event was cancelled due to the COVID-19 pandemic.

Table 3 displays the regression results of Hypothesis 1. As a reminder, Hypothesis 1 states the following. On average, female artists receive worse rankings in the combined vote of the jury and public. The results show that female odds of getting ranked higher (worse) than a male is 1.79 times greater in a model without controls (1), ceteris paribus. The model with controls (2) shows that females have 2.21 times greater odds of getting ranked higher than men, ceteris paribus. Both effects of models 1 and 2 are significant at the 1% level. The pooled OLS model with controls (4) shows that females' odds of getting ranked higher than males are, on average, 2.98, ceteris paribus. The fixed effect model with controls (6) shows that females get ranked on average 3.23 ranks higher than males, ceteris paribus. Again, both effects of pooled OLS and fixed effects are significant at the 1% level. In addition, all control variables in all models in Table 3 also show significant effects at the 1% level. Interestingly, adding control variables to a model always increases the effect of Female on Final Rank.

Table 3: Regression Results Hypothesis 1

	Ordered Logit Year FE		Pooled OLS		FE	
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Female	1.79***	2.21***	2.48***	2.98***	2.83***	3.23***
	(0.09)	(0.13)	(0.22)	(0.23)	(0.26)	(0.25)
Intercept			12.41***	13.58***	12.24***	10.87***
			(0.15)	(0.63)	(0.16)	(0.70)
Controls (English, Order, Solo, Host)	no	yes	no	yes	no	yes
Observations	4,547	4,547	4,547	4,547	4,547	4,547
χ^2	124.32	475.96				
R-squared			0.027	0.081	0.027	0.075

Note: Logit model in odd ratio. Standard errors in parentheses. Significance levels: *: 10%, **: 5%, ***: 1%. All models take Final Rank as the dependent variable. Female equals 1 if an artist's gender is female.

Consequently, all six models reject the null hypothesis that Female is not associated with Final Rank at the 1% level. These results are in line with the first hypothesis and show that, on average, females get ranked worse than males. Thus, results show that the variable Female is associated with higher (worse) ranks. Furthermore, the time series and the regression results show a strong positive effect of being female on the final rank.

Hypothesis 2

As a reminder, Hypothesis 2 states the following. On average, females get ranked worse with decreasing GGGI in both the combined vote of the jury and the public. Figure 3 shows a scatterplot between the Final Rank of a contestant and the GGGI of a voting country. In addition, the scatterplot is split into two groups, Female and Male to show differences based on gender. The variable GGGI indicates a wider gender gap in a country with decreasing values. The confidence interval is visualised in both scatterplots and displays the 5% confidence interval.

When comparing the two scatterplots, one can observe that the difference between the Final Rank of female and male artists decreases with a closer gender gap and increases with a wider gap. This effect can be observed because the slope for females is negative, and the slope for males is positive. The results are in line with the second hypothesis. However, the added value of showing the scatterplot is not obtained by discussing the slopes of a linear model. Rather, the benefit of showing the plot is by observing the confidence interval. As assumed earlier when describing the distribution of the GGGI in Figure 1, both confidence intervals get wider the higher the GGGI. The confidence intervals widen due to the skewed distribution. Since the

extremes are quite important to observe a significant difference in slope gradients, fewer observations among the high GGGI become a problem.

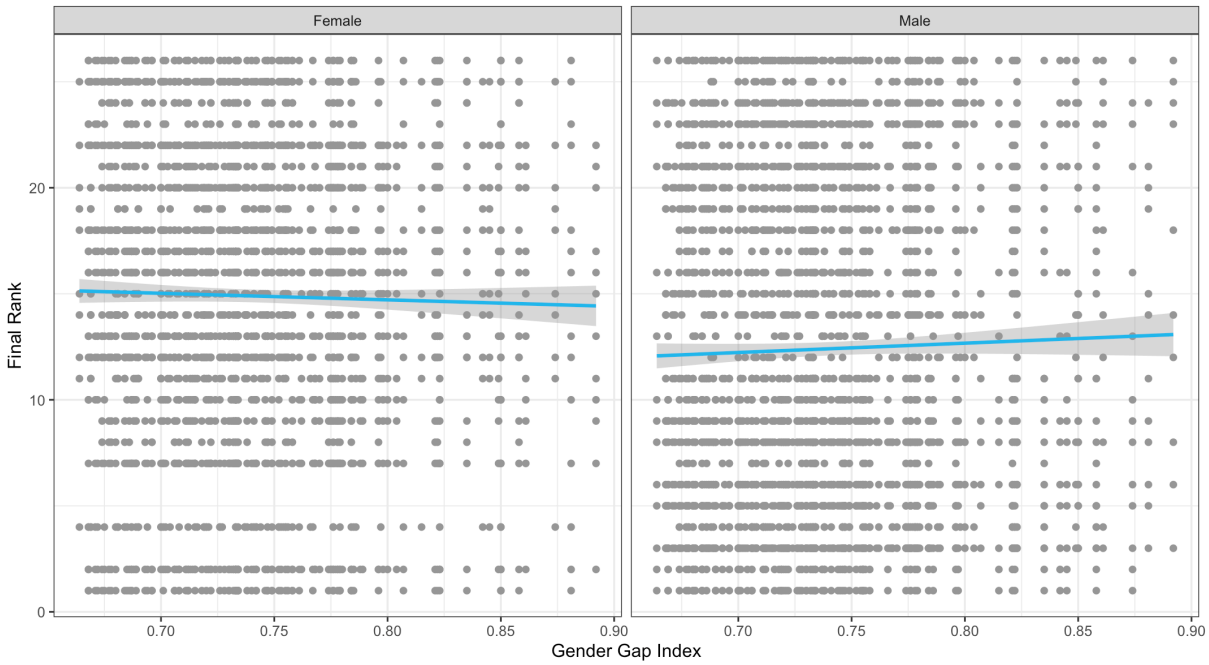

Figure 3: Scatterplot between Final Rank and GGGI Grouped by Gender

Note: The grey area displays 5% confidence intervals.

However, no interpretations about the significant level can be formulated yet. The regression results will allow us to understand better whether the GGGI predicts significant closer or bigger differences between males and females.

Table 4 displays the regression results to test Hypothesis 2. The difference to the first is that an interaction effect between Female and the GGGI is added. The variable Female indicates female artist's with the value 1 and males with the value 0. As mentioned before, the variable GGGI indicates a wider gender gap in a country with decreasing values and vice versa.

Similar to the regression results of the first hypothesis, the ordered logit results show that the odds of a female artist getting ranked higher (worse) than a male are 6.77 times greater in a model without controls (1), ceteris paribus. The effect is significant at the 5% level. The model with controls shows that the odds of females getting ranked higher than males are 7.51 times greater, ceteris paribus. The effect is significant at the 1% level. The effect of being female on the final rank is, on average, plus 8.04 and 8.09 ranks in the pooled OLS model without controls

and with controls, respectively, ceteris paribus. Both effects are significant on the 5% level. The effects of Gender in the FE models are insignificant at the 10% level.

The results of the GGGI are not significant at the 10% level in the ordered logit and pooled OLS model. This result is not surprising because the GGGI of a voting country should not be associated with the variable Final Rank. Every country has to vote for 25 (if participating in the final) or 26 countries. Therefore, the average of Final Rank will always be around 13.5 (25 ranks) or 13 (26 ranks) no matter what the GGGI is. However, we observe significant positive effects of the variable GGGI in the fixed effects model, ceteris paribus. The results of the variable GGGI are not important for the second hypothesis and will not be further discussed.

As mentioned before, the main focus of this regression analysis is the interaction term. We cannot observe significant results at the 10% level in any model with control variables. Nevertheless, the ordered logit model and the pooled OLS model without controls show significant negative effects at the 10% level. The effect can be interpreted in the following way. Females (value 1) get ranked lower (better) with an increasing GGGI score. That means that the results in models 1 and 3 indicate that a wider gender gap correlates with worse rankings of females on the 10% significance level. Furthermore, all results, including those from Figure 3, show directional information in line with the second hypothesis.

Although two results are significant on the 10% level and all directional results are in line with the hypothesis, overall, we cannot state that the results support the second hypothesis. First, the effects are not significant among the models with controls and second, the significance level of the other two models is quite low at 10%. Thus, the null hypothesis that the effect of Gender*GGGI is 0 cannot be rejected.

Table 4: Regression Results Hypothesis 2

	Ordered Logit Year FE		Pooled OLS		FE	
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Female	6.77**	7.51***	8.04**	8.09**	6.24	5.64
	(5.28)	(5.84)	(3.36)	(3.26)	(3.88)	(3.59)
GGGI	2.80	2.25	4.41	4.04	17.40*	15.25*
	(2.10)	(1.68)	(3.15)	(3.06)	(9.07)	(8.40)
Female*GGGI	0.17*	0.19	-7.50*	-6.90	-4,12	-3.27
	(0.17)	(0.20)	(4.52)	(4.39)	(5.21)	(4.82)
Intercept			9.14***	10.59***	-0.65	-0.43
			(2.34)	(2.35)	(6.72)	(6.26)
Controls (English, Order, Solo, Host)	no	yes	no	yes	no	yes
Observations	4,547	4,547	4,547	4,547	4,547	4,547
χ^2	127.31	478.45				
R-squared			0.027	0.082	0.022	0.073

Note: Logit model in odd ratio. Standard errors in parentheses. Significance levels: *: 10%, **: 5%, ***: 1%. All models take Final Rank as the dependent variable. Female equals 1 if an artist's gender is female.

In addition to the regular regressions with the overall GGGI, robustness tests are run. Here the sub-indices of the total GGGI (economic, educational, health and political) are used to test whether certain aspects of gender inequalities are affecting differences in voting behaviour related to gender. Theoretically, one could argue that, for example, gender gaps in health might affect voting differently than differences in political representation. Table 6 in the appendix displays the eight regression results. Each regression is an ordered logit model with year fixed effects. For each sub-index, two regressions are run, one without controls and one with controls. The method does not differ from the first two regressions in Table 4. However, the term "SubGGGI" is a filler for the according sub-index.

All odd ratios of the interaction terms Gender and SubGGGI are between 0.99 and 1.01. Hence, none of the subindices can explain differences in the variable Female. Independent from the significance level, the magnitude of the effects is too small to support the second hypothesis. Neither economic, educational, health, nor political factors can explain differences in the variable Female on a bigger scale.

The robustness test results support the interpretation of the results in Table 2. Neither the overall index nor the subindices can reject the second null hypothesis.

Hypothesis 3

Last but not least, Hypothesis 3 is tested. It states the following. The effect of Female on rank is significantly bigger for Tele Rank compared to Jury Rank. Before interpreting the regression results, once again, a time series is shown in Figure 4. It displays the average rank of females among the tele and jury votes. Given the descriptive statistics shown in Table 1 the average Final Rank in this dataset is 13.6. Consequently, if we assume there is no gender bias, we would expect an average rank of 13.6 for both Jury-Female and Tele-Female Rank. However, this bias-free assumption is not supported. We can see in Figure 4 that the jury is closer to this bias-free benchmark than the tele voters. On average, females get ranked worse by the tele than jury votes. The difference in female ranks between jury and tele vote varies between approximately 2 and 5 ranks. Although there are year effects between males and females (sometimes the gap is smaller or bigger), one can observe a trend. The described trend is in line with the third hypothesis. The observations raise the question of why jury and tele votes differ in such a manner over time. If judgments and evaluations are purely based on the quality of performance, one would not expect much difference in Jury and Tele Rank.

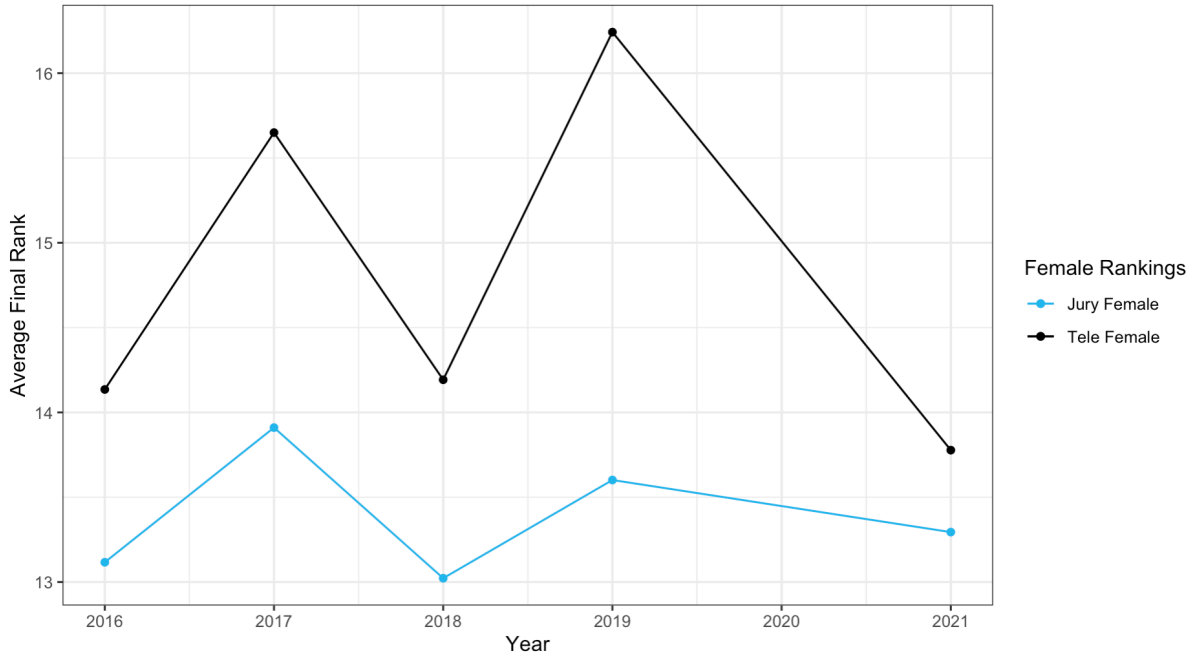

Figure 4: Time Series of Final Rank between Female Jury and Female Tele Vote

Note: 2020 is missing as the event was cancelled due to the COVID-19 pandemic.

To test Hypothesis 3 in a regression approach, differences in jury and tele votes are tested. Now, Final Rank is not the main dependent variable anymore. Instead, Jury or Tele Rank is taken as the main dependent variable. The two sets of ranks are indicated by the variable Jury Rank, which takes the value of 1 if the rank is given by the jury of experts and 0 if the public viewers of a country give the rank. Furthermore, an interaction variable is added to capture differences between jury and tele votes. According to the hypothesis, the effect of the interaction variable between Female and Jury Rank has to be negative in linear models and below 1 in log models. The interpretation of such an effect would be that, on average, jury members rank females lower (better) than tele voters.

The original dataset was adjusted to test the significance and magnitude of the differences between jury and tele vote. In the original voting data, each year has three rank values per observation. Every country that votes for another country in a year produces a Jury, Tele and Final Rank. When Hypothesis 1 and 2 are tested, only Final Rank is of interest. The unique identifier changes when testing differences between Jury and Tele Rank. Now, every year and voting pair has two observations, one for jury rank and one for tele rank. Thus, the observation count doubles compared to the regression results in Tables 3 and 4.

Table 5 displays the result of the six regressions that measure the difference in voting between jury and tele votes. In line with Hypothesis 3 and Figure 4, all six regression results show positive effects for the variable Female, ceteris paribus. The effects are all significant at the 1% level. More importantly, the interaction effects in the ordered logit model show that females' odds of getting ranked higher are 0.59 times lower for jury vote compared to tele vote, ceteris paribus. The effect is significant at the 1% level. The effects of a model with or without controls do not differ. The pooled OLS and fixed effects regression results can be interpreted more intuitively. Pooled OLS results show that, on average, females get ranked 2.25 lower (better) when judged by a jury compared to tele votes, ceteris paribus. The effect is the same for both models with and without controls and is significant at the 1% level. Furthermore, the fixed effects model supports these findings. Here, females get ranked on average 2.32 ranks lower (better) when ranked by a jury compared to getting ranked by tele votes, ceteris paribus. Once again, the effect is significant at the 1% level. All models show results that are in line with the first hypothesis. Therefore, the null hypothesis that the effect of Female on Jury Ranks is not different from the effect of Female on Tele Ranks can be rejected at the 1% level.

Table 5: Regression Results Hypothesis 3

	Ordered Logit Year FE		Pooled OLS		FE	
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Female	1.91***	2.06***	2.75***	2.99***	2.69***	2.82***
	(0.10)	(0.11)	(0.21)	(0.22)	(0.21)	(0.21)
JuryRank	1.22***	1.20***	0.88***	0.88***	1.02***	1.03***
	(0.06)	(0.06)	0.21)	(0.21)	(0.23)	(0.23)
Female*JuryRank	0.59***	0.59***	-2.25***	-2.25***	-2.32***	-2.32***
	(0.04)	(0.04)	(0.30)	(0.30)	(0.29)	(0.29)
Intercept			11.99***	14.28***	11.96***	13.75***
			(0.15)	(0.45)	(0.16)	(0.43)
Controls (English, Order, Solo, Host)	no	yes	no	yes	no	yes
Observations	9,094	9,094	9,094	9,094	9,094	9,094
χ^2	177.33	469.86				
R-squared			0.019	0.048	0.018	0.047

Note: Logit model in odd ratio. Standard errors in parentheses. Significance levels: *: 10%, **: 5%, ***: 1%. All models take Rank as the dependent variable. Female equals 1 if an artist's gender is female.

The results show that gender effects are significantly bigger among tele votes when compared to jury votes. Since the rankings of jury and tele are valued equally, we can conclude that the majority of observed gender effects in Final Rank come from tele voters.

Table 7 in the appendix displays regression results that measure the effect of Female on Jury and Tele Rank separately. For the ordered logit model, the results for jury votes show that being female increases the odds of getting ranked higher than males by 1.2, ceteris paribus. When analysing tele votes, the results show that being female increases the odds of getting ranked higher than males by 2.15, ceteris paribus. In both cases the effects are significant on the 1% level. Interestingly, the fixed effects models show slightly different results. On the one hand, on average, being female increases the Jury Rank by 0.18 ranks, ceteris paribus. However, the effect is not significant at the 10% level. On the other hand, on average, being female increases the Tele Rank by 2.5 ranks, ceteris paribus. The effect is significant at the 1% level. The results support the claim that jury members are less affected by the female of an artist when evaluating their performance.

Based on this finding, another robustness test is run. It tests whether possible interaction effects between the variables Gender and GGGI are significant among tele votes. Table 8 in the appendix displays regression results with an interaction effect between Female and GGGI for the

jury and tele votes separately. Surprisingly, contradicting results to the second hypothesis is observed. First, stronger effects among tele voters are observed, but in unexpected directions. All models show positive directions in the interaction variable. However, with low significance levels, 10% in the ordered logit model, 5% in the OLS model and the FE model are insignificant at the 10% level. Positive directional results mean that the models predict that females get ranked higher with greater GGGI values (gap closes). This result contrasts hypothesis two, which assumed the exact opposite. Nonetheless, the results support the claim (raised when discussing hypothesis two) that the GGGI fails to explain gender effects in the voting behaviour of the ESC.

Discussion

This chapter will discuss the results and their implications. Before discussing results, an economics framework will be defined, followed by a short digression explaining aspects of music preferences. The economic framework will help us understand the dependencies in the model and the challenges when interpreting the regression results. It helps answer questions regarding what factors can be measured, what assumptions must hold to measure gender bias, and whether these assumptions hold in the obtained results. Next, music preferences are discussed to show the complexity of music preferences and evaluations, particularly in popular music. Subjective music preferences might differ greatly. In general, this chapter aims to formulate the limitations of the research, discuss the implications and external validity of the results and suggest recommendations for future research.

Economic Framework

The basic problem faced by each country *j* is to rank each country *i*'s performance. As explained in the chapter where the ESC was introduced, the count of votes for favourite songs given by the public is used to rank countries labelled as tele votes. On the other side, the jury members rank each of the 26 songs.[11] Finally, the average of the five jury members is used to calculate

[11] Note that jury and televoters cannot vote for their own country. Hence, in some cases, 25 songs are ranked instead of 26. However, that does not affect the economic model and its assumptions. In the following, 26 ranks are assumed.

the overall Jury Rank. The basic modelling assumption, including gender bias, is that this decision will depend on the following factors. The first factor is the *perceived quality* of each song. *Perceived quality* can be further decomposed into *objective quality* that relates to qualitative performance and song attributes, *country quality* which is a country's idiosyncratic preferences for a certain type of music, performance or song and *subjective quality* that relates to very personal preferences of each voter.

Another factor and the one of main interest in this research is *gender*. To become more concrete, g_{ij} is the effect of *gender* of a country's *i* artists on a country's *j* ranking. In addition, several other influences and biases influence Final Rank. They are summarised in a factor variable c_{ij}. It captures all influences controlled for in the regression model, including effects of *order*, being *host*, the *language* of a song and *solo* capturing the number of people performing on stage. Lastly, an error term ε_{ij} captures normal noise in the data and possible omitted effects that influence final rank. The quality of country *i*'s entry as perceived by country *j* is denoted as $q_{ij} = \theta_i + \iota_{ij} + s_{ij}$, where θ_i is the objective song attribute, ι_{ij} represents country *j*'s idiosyncratic preference for country *i*'s song and s_{ij} represents the subjective preference of voter in country *j* for country *i*'s song. For example, the speed of a song or choice of the instrument would affect both country's idiosyncratic and subjective preferences. However, idiosyncratic preferences would be a country's baseline, and subjective preferences would be individual differences. One might raise the concern that subjective preferences should not matter since only a country's averages matter. However, the thought is only valid if the voting population represents a country's population.

Consequently, the overall valuation of each performance can be mapped to a one-dimensional index. $v_{ij} = f(g_{ij}, \theta_i, \iota_{ij}, s_{ij}, c_{ij}, \varepsilon_{ij})$. Songs are ranked according to the value of v_{ij}. The highest v_{ij} is ranked first, the second-highest is ranked second, and so on up to the 26th rank. The model shows that we can only measure gender effects independent of perceived quality if we can measure all other factors or if the assumption holds that certain parameters are, on average, equal among countries. On top of that, there is an assumption that the voter's sample represents a country's overall population.

Music preferences

Before interpreting results regarding the effects of gender on voting in a musical contest, one has to think about how gender affects evaluations in the music domain and how preferences and expectations among genders may differ in this regard. These aspects are of particular interest to further understand preferences, differences and possible biases. It is far from certain to conclude that there are discriminating gender biases only because there is a correlational effect. Many possible other factors may be omitted, biasing the results. The subsequent discussion aims to introduce factors related to music preference discussed in the literature.

Elliot (1995) concludes that there is evidence of strong masculine or feminine associations between musical instruments and gender. The research findings seem to indicate that such stereotyping may influence evaluations of musical performances. Interestingly, these effects were only observable for women, not for men. More recent research by Cumberledge (2018) contradicts these findings. The author concludes that although behavioural stereotyping may be present in perceptions made of wind band musicians, males and females were judged equally regarding their musical performance.

However, the evaluation of a musical performance is not only influenced by the quality of a performed song. In modern music, it is common to see dances and perform choreographies. Juchniewicz (2008) analysed three physical movement conditions no movement, head and facial movement and full-body movement in pianists' performances. Results indicate that the pianist's physical movements significantly increased participants' performance ratings. Hence, musical ability affects rating and other factors such as physical movement. One can assume that more complex movements in modern popular songs can also affect performance ratings. On top of that, performances and appearance standards differ depending on gender. Aubrey and Frisby (2011) tried to measure these differences in modern music videos. They compared sexual objections across artists' gender and musical genres. They find suggestive evidence that compared to male artists, female artists were more sexually objectified, held to stricter appearance standards, and were more likely to demonstrate sexually alluring behaviour. It is hard to say what factors affect these differences. Factors such as pressure in the industry, different expectations from the audience or different preferences among female and male artists could matter. Either way, the different "styles" of performances will likely influence the audiences'

perception of quality or their liking of the performance. This effect might lead to an omitted variable bias in this research. On the one side, songs are likely to be unequally performed. Hence, depending on distributions, certain types of performances could be over-represented based on gender. On the other side, audience expectations might depend on gender. This dependency will lead to a bias if gender is not equally distributed among voters. For example, if more males than females vote in the ESC, then male preferences would dominate the overall ranking. Thus, if those preferences are associated with the gender of a performing artist, then gender effects would not measure discrimination but effects created by selection bias.

Furthermore, the attractiveness of artists may influence evaluations as well. On this matter, Ryan and Costa-Giomi (2004) investigated how attractiveness bias may influence the judgment and evaluation of young pianists' performances. Their results show that an attractiveness bias may affect evaluations of audio-visual recordings of musical performances. More attractive pianists were ranked higher. Wapnick et al. (1997) found similar results with the caveat that there were confounding results for females. Comparing audio-alone and audio-visual ratings, more attractive females performed better in both cases. They raise the idea that more attractive people might get more support and encouragement in life, which might improve their ability to learn and train. Ryan et al. (2006) could not find significant effects when testing whether the attractiveness of high-class pianists affects their evaluations. Though, their findings indicate that men and women reacted differently to attractiveness. They raise the question of whether men and women might have different perceptions of what constitutes an attractive performer.

As mentioned before, not only musicians may affect ratings, but also an audience's preferences affect a musician's rating. For example, Delsing et al. (2008) found significant differences in music preferences among genders in the Dutch populations. Overall, males showed stronger preferences for Rock, whereas females showed stronger preferences for Elite and Urban genres. Instead of exploring gender differences in genre preference, Miller (2008) aimed to explore gender differences in artists' preferences by establishing that participants' lists of favourite artists contained an unequal ratio of males and females. Consistent with prior research, men made up the majority.

The discussion about music preferences could be continued. However, the purpose was to show the complexity of music preferences and their effects on subjective preferences, measured in

the economic model as s_{ij}. It becomes clear that the assumption that subjective preferences are, on average equal may not hold, especially if selection bias is present. Next, the implications of the observed results are discussed.

Discussion regarding obtained results

According to the three hypotheses, the regression results show three key findings. First, the results show significant gender effects in the voting data. According to the regression results in Table 2, in 2016 to 2021, females rank on average 2.5 to 3 below males. The observed gender effects are in line with prior research in the musical industry that reported anti-female gender bias (Colley and North, 2003; Golding and Rouse, 2000). Second, there is no significant association between the GGGI of a country and the effect of gender. Thus, the results could not show that gender inequality in societies affects the evaluation of artists in these societies based on their gender. Third, jury and tele votes differ significantly. Moreover, average Jury Ranks were closer to the expected rank of 13.6 than Tele Ranks indicating a more objective vote. More objective votes by the jury support findings of Haan et al. (2005), who conclude that in the ESC, experts are less sensitive to factors unrelated to quality than the public. Therefore, the European Broadcasting Union should consider only using Jury Ranks to determine the winner if the contest's goal is to provide a fair vote for the best musical performance.

Hence, on the one side, there are observed effects in g_{ij} that are in line with the hypothesis that such gender effects would be significantly greater among public votes compared to jury votes. On the other side, the GGGI fails to explain these effects and thus fails to support a claim of gender bias. Therefore, comparing jury and public votes raises the question of where the difference originates, if not from gender gaps in certain countries. There are several possible answers to this question.

First, there may be discriminating gender bias stronger among public votes than jury votes, but the GGGI does not significantly explain them. There may be gender bias in music evaluations that are not associated with GGGI of an evaluating country. However, this research failed to find results that support this point.

Second, present selection bias may be a reason for gender effects in voting. Selection bias, which means a non-random selection of voters, would violate the economic model's assumption and bias the results. Following the economic model, perceived quality ($q_{ij} = \theta_i + \iota_{ij} + s_{ij}$) should be on average even when comparing genders. That is because different subjective preferences based on gender (as discussed above) should equalise if gender among voters and musicians is evenly distributed. Although we observe evenly distributed genders among artists (48% Females), this might not be the case for voters. If the ESC contest, and more importantly, the voting itself, attracts more males than females or vice versa, subjective preference (s_{ij}) would not be on average even. That is because different genders have, on average different music preferences. As discussed before, the attractiveness of performing artists, different music genres or choreographies have been observed to differ among genders. On top of that, selection must not only be based on gender. Certain people may be attracted to the contest because they like certain performances or genres independent of gender. This selection may happen in the voter's base. On the other side, juries get elected and constructed by the organisers of the ESC. Therefore, juries are not affected by a "free" selection bias and rather get selected in a controlled way. This difference may be a driver in differences when judging the music. As discussed above, females and males do represent different music genres and are associated with different styles. This difference may lead to gender effects that do not measure discrimination but rather proxies for legitimate music preferences. Thus, gender bias cannot be inferred from current data due to uncertainty about the voter population.

This point leads to a recommendation for future research to measure whether viewers of the ESC and, more particular, voters are a representative sample of a country's population. Further, it would be interesting to analyse if there is a difference in voting behaviour, including effects of gender, between the voting base of the ESC and a randomly selected sample. Gender effects still occurring in a randomly selected sample would provide suggestive evidence for gender bias.

Lastly, another big difference between jury and tele votes is their voting methods. First, public decision-making is secretive, whereas the jury votes are transparent.[12] Levy (2007) shows that when the decision-making process is secretive (when individual votes are not revealed to the public), committee members are more prone to biases. Hence, since individual jury votes are revealed to the public, and jury members are named in public, jury members might fear negative reputation effects causing less biased voting. Future research may analyse this effect in settings of the contest. One could conduct an experiment mimicking a music contest. Two groups of experts or the public would vote, one public and the other secretive. Different kinds of biases could be tested and evaluated. Second, juries rank all performances once. Furthermore, tele votes are asked to vote for their favourite song. They may vote multiple times. In the end, the performance with the highest count ranks first. Therefore, tele voters do not rank all performances and are allowed to increase their vote count by voting multiple times for the same contestant. This makes comparing Jury and Tele Rank challenging. Different rules and setups may lead to different outcomes independent of their biases towards the gender of an artist. Because of this ambiguity, gender bias cannot be inferred from the different results in jury and tele votes.

External validity is given to a limited extent due to the discussed limitations. Differences in voting rules and selection processes between jury and televoters lead to inaccurate results. Furthermore, only correlational effects were observed. Nonetheless, the significant differences between jury and tele vote support the findings of Haan et al. (2005) and Cleredes Stengos (2012). Since jury votes are closer to the overall average Final Rank of 13.6, we can assume that jury votes are more objective considering gender effects than tele votes. This effect probably also applies outside the ESC context, supporting the claim that expert votes are more appropriate when evaluating performances.

[12] Note that individual jury members are labelled A, B, C, D, and E in the voting dataset. Thus, the voting is not fully transparent. However, unlike tele voters, jury members are named publicly and know ex-ante that their results will be tracked and published.

Conclusion

In conclusion, the regression results discussed in this work show gender effects in the ESC voting data between 2015 and 2021. On average, women got ranked significantly worse than men. Furthermore, the effect is significantly stronger among tele votes than jury votes. In contrast to the second hypothesis, the GGGI failed to explain differences in gender effects. Thus, the results could not show that gender inequality in societies affects the evaluation of artists in these societies based on their gender. Moreover, several limitations have to be considered when interpreting the results. In particular, a self-selecting voting base among the ESC viewers could have led to selection bias in the voting data. Therefore, it is not possible to conclude that causal effects were observed when measuring gender effects, nor is it possible to conclude that artists are affected by discriminating gender bias in the ESC. Nevertheless, the results support prior research that suggests jury members are less prone to biases and will vote more objectively.

References

Aubrey, J. S., and Frisby, C. M. (2011). Sexual objectification in music videos: A content analysis comparing gender and genre. *Mass Communication and Society*, 14(4), 475-501.

Blangiardo, M., and Baio, G. (2014). Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models. *Journal of Applied Statistics*, 41(10), 2312-2322.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral?. *The Quarterly Journal of Economics*, 135(1), 269-327.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

Coupe, T., Gergaud, O., and Noury, A. (2018). Biases and Strategic Behaviour in Performance Evaluation: The Case of the FIFA's best soccer player award. *Oxford Bulletin of Economics and Statistics*, 80(2), 358-379.

Clerides, S., and Stengos, T. (2012). Love thy neighbor, love thy kin: Strategy and bias in the Eurovision Song Contest. *Ekonomia*, 15(1).

Colley, A., North, A., and Hargreaves, D. J. (2003). Gender bias in the evaluation of New Age music. *Scandinavian Journal of Psychology*, 44(2), 125-131.

Crotti, R., Pal, K. K., Ratcheva, V. and Zahidi, S. (2021). The global gender gap report 2021. *World Economic Forum*.

Dekker, A. (2007). The Eurovision Song Contest as a ‘friendship’ network. *Connections*, 27(3), 53-58.

Delsing, M. J., Ter Bogt, T. F., Engels, R. C., and Meeus, W. H. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality*, 22(2), 109-130.

European Broadcasting Union (n.d. a). How it works. https://eurovision.tv/about/how-it-works

European Broadcasting Union (n.d. b). In a Nutshell. https://eurovision.tv/history/in-a-nutshell

European Broadcasting Union (2021). 183 million viewers welcome back the Eurovision song contest. https://eurovision.tv/story/183-million-viewers-welcome-back-the-eurovision-song-contest

European Broadcasting Union (2022a). Rules. https://eurovision.tv/about/rules

European Broadcasting Union. (2022b). History. https://eurovision.tv/events

Elliott, C. A. (1995). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education*, 50-56.

Fehr, E., and A. Rangel (2011), Neuroeconomic Foundations of Economic Choice. Recent Advances. *Journal of Economic Perspectives*, 25, 3-30.

Fenn, D., Suleman, O., Efstathiou, J., and Johnson, N. F. (2006). How does Europe make its mind up? Connections, cliques, and compatibility between countries in the Eurovision Song Contest. *Physica A: Statistical Mechanics and its Applications*, 360(2), 576-598.

Flores, R., and V. Ginsburgh (1996), The Queen Elisabeth musical competition: How fair is the final ranking, *Journal of the Royal Statistical Society*, Series D, 45, 97-104.

Fryer Jr, R. G., and Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2), 210-40.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice*, 4(2), 99.

Gatherer, D. (2006). Comparison of Eurovision Song Contest simulation with actual results reveals shifting patterns of collusive voting alliances. *Journal of Artificial Societies and Social Simulation*, 9(2).

Gibbard, A. (1973). Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, 587-601.

Ginsburgh, V., and Moreno-Ternero, J. D. (2022). The Eurovision Song Contest: Voting Rules, Biases and Rationality. *ECARES Working Papers*, 2022.

Ginsburgh, V., and Noury, A. G. (2008). The Eurovision song contest. Is voting political or cultural?. *European Journal of Political Economy*, 24(1), 41-52.

Glejser, H., and Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth Music Contest. *Journal of Cultural Economics*, 25(2), 109-129.

Goldin, C., and Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, 90(4), 715-741.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science*, 320(5880), 1164-1165.

Haan, M. A., Dijkstra, S. G., and Dijkstra, P. T. (2005). Expert judgment versus public opinion–evidence from the Eurovision song contest. *Journal of Cultural Economics*, 29(1), 59-78.

Hamberg, K. (2008). Gender bias in medicine. *Women's health*, 4(3), 237-243.

Harris, M. N., Novarese, M., and Wilson, C. M. (2022). Being in the right place: A natural field experiment on the causes of position effects in individual choice. *Journal of Economic Behavior and Organization*, 194, 24-40.

Herr, P. M., Sherman, S. J., and Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of experimental social psychology*, 19(4), 323-340.

Juchniewicz, J. (2008). The influence of physical movement on the perception of musical performance. *Psychology of Music*, 36(4), 417–427.

Kelman, M., Rottenstreich, Y., and Tversky, A. (1996). Context-dependence in legal decision making. *The Journal of Legal Studies*, 25(2), 287-318.

Kocher, M.G., and M. Sutter (2006), Time is money. Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior and Organization*, 61, 375-392.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., and Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin*, 137(4), 616.

Levy, G. (2007). Decision making in committees: Transparency, reputation, and voting rules. *American economic review*, 97(1), 150-168.

Long, J. S., and Freese, J. (2006). Regression models for categorical dependent variables using Stata (Vol. 7). *Stata press*.

Lynch Jr, J. G., Chakravarti, D., and Mitra, A. (1991). Contrast effects in consumer judgments: Changes in mental representations or in the anchoring of rating scales?. *Journal of Consumer Research*, 18(3), 284-297.

McKelvey, R. D., and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, 4(1), 103-120.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2), 109-127.

Rand, D. G., Pfeiffer, T., Dreber, A., Sheketoff, R. W., Wernerfelt, N. C., and Benkler, Y. (2009). Dynamic remodeling of in-group bias during the 2008 presidential election. *Proceedings of the National Academy of Sciences*, 106(15), 6187-6191.

Rudman, L. A., and Glick, P. (2021). The social psychology of gender: How power and intimacy shape gender relations. *Guilford Publications*.

Ryan, C., and Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education,* 52(2), 141-154.

Ryan, C., Wapnick, J., Lacaille, N., and Darrow, A.-A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of Music*, 34(4), 559–572.

Robson, C., and Walsh, B. (1974). The importance of positional voting bias in the Irish general election of 1973. *Political Studies*, 22(2), 191-203.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2), 187-217.

Senaviratna, N. A. M. R., & Cooray, T. M. J. A. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 5(2), 1-9.

Sigelman, L., and Sigelman, C. K. (1982). Sexism, racism, and ageism in voting behavior: An experimental analysis. *Social Psychology Quarterly*, 263-269.

Spierdijk, L., & Vellekoop, M. (2009). The structure of bias in peer voting systems: lessons from the Eurovision Song Contest. *Empirical Economics*, 36(2), 403-425.

Stockemer, D., Blais, A., Kostelka, F., and Chhim, C. (2018). Voting in the Eurovision Song Contest. *Politics*, 38(4), 428–442.

Tversky, A., and Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39(10), 1179-1189

World Economic Forum. (2022). Global Gender Gap Report. https://www.weforum.org/search?query=global+gender+gap+report

Appendix

Table 6: Regression Results with Sub-Indices – Hypothesis 2

	Ordered Logit Model Year FE (Economic SubGGG)		Ordered Logit Model Year FE (Educational SubGGG)		Ordered Logit Model Year FE (Health SubGGG)		Ordered Logit Model Year FE (Political SubGGG)	
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	2.02***	2.49***	1.98***	2.39***	1.41***	1.83***	2.03***	2.46***
	(0.15)	(0.19)	(0.30)	(0.36)	(0.16)	(0.21)	(0.14)	(0.18)
SubGGGI	1.001*	1.00	1.00	1.00	0.99*	1.00*	1.00*	1.00
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Female* SubGGGI	1.00*	1.00*	0.99	1.00	1.01*	1.01*	1.00**	1.00*
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Controls (English, Order, Solo, Host)	no	yes	no	yes	no	yes	no	yes
Observations	4,547	4,547	4,547	4,547	4,547	4,547	4,547	4,547
χ^2	252.58	955.42	249.19	952.25	253.73	955.04	253.31	954.52

Note: Logit model in odd ratio. Standard errors in parentheses. Significance levels: *: 10%, **: 5%, ***: 1%. All models take Final Rank as the dependent variable. Female equals 1 if an artist's gender is female.

Table 7: Effect of Female on Jury and Tele Ranks

	Ordered Logit Year FE		Pooled OLS		FE	
	Jury	Tele	Jury	Tele	Jury	Tele
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Female	1.20***	2.15***	0.78***	2.95***	0.18	2.52***
	(0.07)	(0.12)	(0.23)	(0.21)	(0.26)	(0.22)
Intercept			14.78***	14.66***	13.81***	14.41***
			(0.63)	(0.60)	(0.72)	(0.62)
Controls (English, Order, Solo, Host)	yes	yes	yes	yes	yes	yes
Observations	4,547	4,547	4,547	4,547	4,547	4,547
χ^2	90.36	506.56				
R-Squared			0.02	0.10	0.02	0.10

Note: Logit model in odd ratio. Standard errors in parentheses. Significance levels: *: 10%, **: 5%, ***: 1%. Jury and Tele indicate the dependent variable Jury Rank or Tele Rank. Female equals 1 if an artist's gender is female.

Table 8: Jury versus Tele Votes with Interaction Effects – Hypothesis 2

	Ordered Logit Year FE		Pooled OLS		FE	
	Jury	Tele	Jury	Tele	Jury	Tele
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.62	0.58	-2.07	-3.20	6.13*	1.23
	(0.49)	(0.45)	(3.27)	(3.10)	(3.69)	(3.18)
GGGI	0.45	0.31	-2.64	-4.75	-8.85	0.80
	(0.34)	(0.23)	(3.07)	(2.91)	(10.32)	(8.90)
Female*GGGI	2.41	5.82*	3.85	8.30**	-8.08	1.89
	(2.52)	(6.11)	(4.40)	(4.17)	(4.97)	(4.28)
Intercept			16.74***	18.18***	20.32***	13.19**
			(2.36)	(2.23)	(7.66)	(6.60)
Controls (English, Order, Solo, Host)	yes	yes	yes	yes	yes	yes
Observations	4,547	4,547	4,547	4,547	4,547	4,547
χ^2	91.50	509.74				
R-Squared			0.02	0.10	0.01	0.10

Note: Logit model in odd ratio. Standard errors in parentheses. Significance levels: *: 10%, **: 5%, ***: 1%. Jury and Tele indicate the dependent variable Jury Rank or Tele Rank. Female equals 1 if an artist's gender is female.