COMPETING IDENTITIES: A FIELD STUDY OF IN-GROUP BIAS AMONG PROFESSIONAL EVALUATORS *

[Short title: COMPETING IDENTITIES]

Anna Sandberg

I use data from the Olympic sport of dressage to explore in-group biases among judges. Dressage – the only international sport with subjective performance evaluations in which men and women compete as equals – provides a rare opportunity to identify multiple in-group biases in the same naturally occurring setting. While, on average, judges are not biased in favour of either gender, they exhibit substantial biases in favour of (i) athletes of their own nationality, and (ii) athletes of the same nationality as the other judges in the competition. Heterogeneity across competitions suggests that biases increase as group identity becomes more salient. (*JEL* J15, J16, J71, L83)

In-group bias, i.e. favouritism of members of one's own group, has important labour market implications. For instance, it may cause discrimination in hiring and promotion decisions and affect team cooperation, judges' verdicts in trials and teachers' evaluations of students. Previous studies demonstrate that individuals often favour members of their own group over members of other groups (e.g. Tajfel *et al*., 1971; Bernhard *et al*., 2006; Chen and Li, 2009; Goette *et al*., 2012). While most previous studies focus on investigating one group at a time, in reality individuals belong to many groups (such as gender and nationality) simultaneously.[1] Thus, it is important to incorporate multiple group identities in the empirical research on in-group bias and explore how different biases interact to reinforce or counteract each other. Moreover, groups may be more or less permanent. While gender and nationality represent

*Corresponding author: Anna Sandberg, Institute for International Economic Studies (IIES), Stockholm University, SE-10691, Stockholm, Sweden. Email: anna.sandberg@iies.su.se.

I thank Hans-Joachim Voth (the editor) and four anonymous reviewers for constructive comments and suggestions that much improved the paper. I am also grateful for helpful comments and advice from Manuel Bagues, Alexander Cappelen, Fredrik Carlsson, Anna Dreber Almenberg, Tore Ellingsen, Karin Hederos, Magnus Johannesson, Matthew Lindquist, Erik Lindqvist, Astri Muren, Frederik Schwerter, David Strömberg, Bertil Tungodden and Robert Östling. I thank the FEI Hedquarters for their help assembling the data set and answering questions, and Isak Wiström for his excellent work collecting the data. Finally, I gratefully acknowledge financial support from the Jan Wallander and Tom Hedelius Foundation.

[1] Whereas previous literature almost exclusively focuses on one in-group bias per setting, two notable exceptions are Dee (2005) and Feld *et al*. (2016), exploring multiple in-group biases in teachers' evaluations of students.

relatively permanent groups, other groups, such as hiring committees, school classes and work teams, tend to be temporary. In this study I use unique data from the equestrian sport dressage to explore in-group biases based on gender, nationality and temporary social groups. Dressage, the only Olympic sport with subjective performance evaluations in which men and women compete as equals, allows clean identification of these in-group biases in the same context, using data on repeated high-stakes decisions of professional decision makers.

There is a large experimental literature on in-group bias.[2] However, endogeneity and limited data availability make it difficult to study this topic in the field. Dressage competitions provide an opportunity to study in-group bias, avoiding the three major identification problems normally associated with naturally occurring data:

(i) Endogeneity. The main obstacle to identifying in-group bias outside the laboratory is the potential correlation between the group membership of evaluators and the relative quality of candidates of different groups. In dressage competitions, each athlete is scored by five judges. Thus, I can identify in-group bias by comparing the scores from five different judges who observed *the same* performance by an athlete, interacting the gender (nationality) of the judge with the gender (nationality) of the athlete. Since each judge is represented multiple times in the data, I can control for average differences in leniency between judges from different groups.

(ii) Data availability. A second widespread obstacle is lack of information on the group membership of evaluators or candidates. For instance, in Goldin and Rouse's (2000) study on gender discrimination in orchestra auditions, lack of information on the jury members prevents the authors from examining how gender discrimination varies with the

[2] Previous experimental studies demonstrate that in-group bias can arise based on natural social groupings such as ethnicity, political affiliation and social networks (e.g. Bernhard *et al*., 2006; Leider *et al*., 2009; Rand *et al*., 2009; Goette *et al*., 2012), and based on more trivial identities induced in the laboratory (e.g. Chen and Li, 2009; Hargreaves Heap and Zizzo, 2009; Sutter, 2009). Numerous studies in social psychology use the minimal group paradigm, assigning subjects to groups based on arbitrary labels. Studies find that even when group assignment is based on seemingly trivial labels, such as preferences over paintings, subjects favour in-group members. The minimal group paradigm was first used by Tajfel *et al.* (1971). See e.g. Bourhis and Gagnon (2001) for an overview of this literature.

gender composition of the evaluators (Bagues and Esteve-Volart, 2010). Moreover, even when data is available on the background characteristics of the committee members, data is often lacking on each member's individual opinion (e.g. Bagues and Esteve-Volart, 2010; Price and Wolfers, 2010). The dressage data include information on the gender and nationality of all athletes and judges. The judges provide scores independently of each other, and the data include each judge's individual scores.

(iii) Lack of diversity among evaluators. In previous studies on gender discrimination, the lack of female evaluators often makes it difficult to identify the interaction between evaluator gender and candidate gender. For example, the small number of female reviewers prevents Wennerås and Wold (1997) from analysing gender interactions when studying gender discrimination in peer-review scores for postdoctoral fellowships.[3] In dressage competitions, there are many judges and athletes of both genders, and most athletes are evaluated by a judging panel containing both male and female judges.

An additional advantage of dressage is that the large number of judges, athletes and competitions make it possible to address how contextual factors and judges' background characteristics affect the strength of biases. Moreover, a new panel of judges is created for every competition. Thus, I can explore the effect of temporary group identity by identifying how a judge is influenced by the other members of the judging panel.

I find that dressage judges favour athletes of their own nationality. The size of the average nationalistic bias is non-negligible, corresponding to 7.2% of the overall standard deviation and 23.8% of the within-performance standard deviation of scores. This bias is particularly striking given that these judges are trained and experienced experts in judging, facing a high degree of monitoring – competitions are observed by audiences, often televised, and all scores are saved in a public database. I presume that a bias needs to be rather strong

[3]Another example is a study by Blank (1991), exploring the effect of double-blind vs. single-blind peer reviewing in the American Economic Review. In this case, the number of female referees reviewing papers written by female authors is too small to analyse the interaction between referee gender and author gender (Bagues and Esteve-Volart, 2010).

and resilient to penetrate this group of evaluators. However, I also find variation across different types of competitions, indicating that the nationalistic bias is in fact malleable. In particular, the bias is largest in competitions in which athletes represent their nation (i.e. championships and team events). This is in line with previous studies demonstrating that in-group bias is affected by the salience of group membership (e.g. Shayo and Zussman, 2011).

On average, I find no gender bias.[4] While some judges favour athletes of their own gender, and others favour athletes of the opposite gender, most judges exhibit no significant bias in either direction. I find some evidence that judges favour athletes of their own gender in competitions that only include a small number of different athlete nationalities. This may indicate that national identity is more salient than gender identity in international settings, causing nationalistic bias to dominate any underlying gender bias in the majority of competitions. I speculate that, in international settings, nationalistic bias may dominate other types of biases as well. If so, recent findings of own-ethnicity bias in national settings (e.g. Price and Wolfers, 2010; Parsons *et al.*, 2011), might not generalize to international settings.

Finally, in addition to favouring athletes of their own nationality, judges favour athletes of the same nationality as the other judges on the judging panel. I show that this pattern is not likely to reflect conscious, strategic behaviour such as vote trading within the panel. Rather, I suggest that the judging panel creates a temporary group identity. Thus, judges exhibit in-group biases based on both a permanent group identity (their own nationality) and a temporary group identity (the nationality of their fellow judges on the panel).

This research adds to a growing literature on in-group favouritism based on nationality and gender among professional evaluators. Previous studies demonstrate that evaluators favour candidates of their own nationality or ethnicity in judicial settings (Shayo

[4] Throughout this paper, I denote the interaction between the judge's gender and the athlete's gender as 'gender bias'. Thus, a positive gender bias (or an own-gender bias) implies that judges favour athletes of the same gender. A negative gender bias (or an opposite-gender bias) implies that judges favour athletes of the opposite gender.

and Zussman, 2011), educational settings (Dee, 2005; Feld *et al*., 2016), financial settings (Fisman *et al.*, forthcoming) and sports settings (Zitzewitz, 2006; Emerson *et al*., 2009; Price and Wolfers, 2010; Parsons *et al*.,2011; and Pope and Pope, 2015). For instance, own-ethnicity bias has been identified among American basketball referees (Price and Wolfers, 2010) and baseball umpires (Parsons *et al*., 2011). The analysis in the current paper most closely resembles that of Zitzewitz (2006), who finds that judges in figure skating and ski jumping favour athletes of their own nationality.

While a positive nationalistic bias is an intuitive prediction, the prediction for gender is more ambiguous. Do evaluators favour others who are similar to themselves also in terms of gender? Or do other factors, such as sexual attraction, lead to favouritism of the opposite gender? So far, the empirical evidence is mixed. To explore the interaction between evaluator gender and candidate gender, previous studies use data from hiring committees (Bagues and Esteve-Volart, 2010; Booth and Leigh, 2010; De Paola and Scoppa, 2015; Bagues *et al*., forthcoming), academic peer review processes (Broder, 1993; Li, 2011; Abrevaya and Hamermesh, 2012), financial settings (Beck *et al*., 2014) and educational settings (Dee, 2005; Breda and Hillion, 2016; Feld *et al*., 2016; Mengel *et al.* 2016; Boring, 2017). While some studies find that evaluators favour individuals of their own gender, other studies find a bias in the opposite direction, no bias, or that the degree and direction of biases depend on situational factors.

To my knowledge, the only previous study that explores favouritism based on both nationality and gender in the same setting is by Feld *et al*. (2016).[5] They randomly assign graders at a Dutch university to exams that did, or did not, contain student names, and find that graders favour students of their own nationality (German vs. Dutch), but not students of

[5] Relatedly, Dee (2005) explores in-group bias based on gender and ethnicity among teachers. He finds that students are more likely to be seen as disruptive by teachers of a different ethnicity or gender. However, as the author points out, he cannot separate active teacher effects (teachers being biased in favour of in-group students) from passive effects (students behaving differently with in-group teachers). In contrast, since I compare judges who observe *the same* performance, any effect I observe will be active (judges providing biased evaluations) rather than passive (athletes adjusting their performances).

their own gender. Relative to this study, I add to the literature by using data that include a large variety of evaluator nationalities, repeated decisions from each evaluator, and temporary as well as permanent groups.[6]

The next Section summarizes the rules of dressage and reports descriptive statistics. Section 2 introduces the main empirical strategy, Section 3 presents the baseline results, and Section 4 explores heterogeneity. Section 5 discusses different behavioural interpretations of the nationalistic bias, Section 6 analyses how judges are influenced by the nationality of the other judges on the panel, and Section 7 concludes.

1. Data

1.1. Dressage

Dressage is the only Olympic sport with subjective performance evaluations in which men and women compete as equals. The sport has been part of the Olympics for more than 100 years, and is sometimes likened to figure skating on horseback.

During a dressage competition each rider performs a series of movements on their horse, one rider at a time. Five judges are placed around the arena, assigning all movements a technical score between 0 and 10.[7] Detailed guidelines regulate how the judges should score each type of movement. The judges also award technical scores for general attributes such as the overall quality of the rider and the horse. A rider's final score is the weighted average of the scores from the five judges, converted to a percentage. Higher weight is given to more difficult movements. To promote consistency in scoring, whenever the scores for a performance differ by more than 5% between the judges they must partake in an evaluation meeting after the competition.

[6] Another advantage of my data is that dressage judges are fully informed about the gender and nationality of all athletes. In the study by Feld *et al.* (2016), the graders infer the nationality and gender of the student through the name on the exam. Thus, as pointed out by the authors, it is likely that some graders (i) confuse Dutch and German names in the non-blind treatment, and (ii) recognize the student's gender from their handwriting in the blind treatment. The later issue may have attenuated the estimated gender bias.

[7] Lower level competitions sometimes include three or four judges, and since 2011 the major championships include seven judges. In my data, 96.7% of performances are scored by five judges.

The highest competition level, the Grand Prix level, includes three types of competitions. In the "Grand Prix" and "Grand Prix Special" competitions, all riders execute the same movements in a pre-determined order, and the judges give technical scores for the precision of the movements. In the "Freestyle to Music" competitions, riders choose their own music and perform individually choreographed patterns of movements. In these competitions, in addition to the technical scores, judges give scores for the artistic quality of the routine.

The governing body of international dressage is the International Federation for Equestrian Sports (FEI). The FEI administers and monitors all international Grand Prix level competitions and manages the licensing and promotion of international judges. Becoming an international judge requires many years of extensive training, including international courses and mentorship from more experienced judges. The FEI appoints the judging panel for championships and other major events. In regular Grand Prix competitions, the judges are appointed by the local organizers (in accordance with FEI rules). Online Appendix D provides more details about the rules governing the appointment and promotion of judges. Overall, the national equestrian federations have relatively little say in the career advancement and appointment of international judges.

The Codex for FEI Dressage Judges (summarized in online Appendix D) states that nationalistic judging is strictly prohibited and subject to sanctions. In addition, judges are required to avoid any conflict of interest such as being the trainer of an athlete or the owner of a horse that takes part in the competition.

1.2. Descriptive Statistics

The data were collected from the FEI website and covers all international Grand Prix competitions taking place between January 2007 and March 2012 as well as the 2006 World

Championships.[8] I summarize the structure of the dataset in Table 1. The data include 90,626 scores for 18,201 performances in 1,533 competitions. Two thirds of performances are by female riders. In total, 1,215 riders from 61 countries and 191 judges from 42 countries are represented in the data.

[Tables 1 and 2 about here]

As shown in Table 2 and Figure A1 in the Appendix, the average scores for male riders and judges are slightly higher than for female riders and judges. This is because, on average, male riders and judges compete and judge on a slightly higher level. On average, judges are considerably older than riders (59 vs. 39 years old).

The World Cup consists of four leagues: the Western European, the Central European, the North American and the Pacific. The Western European League includes the best riders, and the majority of observations in the data (65%) are from competitions in Western Europe. In Figures A2 and A3 in the Appendix, I illustrate the sample size by judge and rider nationality. Germany and the United States provide the largest number of riders (153 German and 161 American riders) and judges (20 German and 15 American judges).

2. Estimation

The key empirical challenge when studying in-group bias in dressage judging is that there is no objective measure of the quality of a performance. This situation is similar to many labour market settings, where it is difficult to objectively quantify the quality of the candidates that are being evaluated. My main identification strategy is to, for each performance, compare the average score from the in-group judges to the average score from the out-group judges. Thus, I compare the scores from different judges who observed *the same* performance, estimating the following model:

[8] The scores from the 2006 World Championships were retrieved from the website www.dressagedirect.com. All other scores were retrieved from the FEI website https://data.fei.org/Calendar/Search.aspx in April 2012.

$$s_{jp} = \alpha \cdot J\&R\ same\ gender + \beta \cdot J\&R\ same\ nationality + \theta_p + \gamma_j + e_{jp}, \quad (1)$$

where the dependent variable is the score given by judge j for performance p, θ_p denotes performance fixed effects and γ_j denotes judge fixed effects. The explanatory variables of interest are the indicator variables *J&R same gender*, taking the value 1 if the judge and the rider are of the same gender, and *J&R same nationality*, taking the value 1 if they are of the same nationality. The average gender (nationalistic) bias is given by the coefficient α (β). A positive coefficient implies bias in favour of the in-group, while a negative coefficient implies bias in favour of the out-group.[9] I use the technical, and not the artistic, score as outcome variable since all performances in the data include a technical score, while only 24% include an artistic score.[10]

I also analyse the data at the level of the judge, and estimate judge-specific biases for the 191 judges in the sample. To identify the judge-specific nationalistic bias β_k, I run a separate regression for each judge *k*:

$$s_{jp} = \beta_k \cdot J\&R\ same\ nationality \cdot I(j = k) + \alpha \cdot J\&R\ same\ gender + \theta_p + \gamma_j + e_{jp}, \quad (2)$$

where $I(j = k)$ is an indicator variable for judge *k*. The coefficient β_k indicates how much, on average, judge *k* deviates from the other judges on the panel when the rider is of the same nationality as judge *k* minus how much, on average, judge *k* deviates from the other judges when the rider is of another nationality. Similarly, the following equation identifies the judge-specific gender bias α_k:

$$s_{jp} = \alpha_k \cdot J\&R\ same\ gender\ \cdot I(j = k) + \beta \cdot J\&R\ same\ nationality + \theta_p + \gamma_j + e_{jp}. \quad (3)$$

[9] Since the hypothesized direction of the bias is less straightforward for gender than for nationality, it is important to keep in mind that, according to this notation, a positive gender bias implies an own-gender bias while a negative gender bias implies an opposite-gender bias.

[10] The main results are robust to using the artistic instead of the technical score as outcome variable.

3. Baseline Results

As shown in the first column of Table 3, the estimated nationalistic bias is 0.359 and statistically significant ($p<0.01$). Thus, a rider receives on average 0.359 more points from judges of their own nationality than from judges of other nationalities. The size of this bias corresponds to 7.2% of the overall standard deviation and 23.8% of the within-performance standard deviation of technical scores. To approximate the effect on a rider's final position in a competition, I deduct 0.359 from all scores from judges of the same nationality as the rider, and compute new positions based on these "bias-adjusted" scores.[11] As a result, the final position changes for 5.0% of all performances, and for at least one rider in 23.9% of all competitions.

[Table 3 about here]

The estimated gender bias is small and statistically insignificant. On average, a rider receives 0.010 more points from judges of the same gender than from judges of the opposite gender. This amounts to 0.2% of the overall standard deviation and 0.7% of the within-performance standard deviation of technical scores. The gender bias is relatively tightly estimated around zero, with a 95% confidence interval of [-0.024;0.043].

The above analyses show that, *on average*, judges exhibit positive nationalistic bias but no gender bias. However, these averages may conceal important heterogeneity between judges. Thus, as the next step, I analyse the data at the level of the judge. In Figure 1, I show the judge-specific results from estimating models (2) and (3). A majority of judges exhibit a positive and statistically significant bias in favour of riders of their own nationality; 83% of the judge-specific coefficients β_k are positive, and 55% are positive and statistically significant ($p<0.05$). Only 4% of the coefficients are negative and significant. Thus, the positive nationalistic bias is not driven by a few extreme judges.

[11] In the "Grand Prix Special" competitions, both the technical and the artistic scores determine a rider's final position. However, for simplicity, I base these calculations on the ranking of riders according to their technical scores only.

The judge-specific gender bias coefficients α_k are centred around zero and the majority are statistically insignificant. However, as illustrated by Figure B1 in the online Appendix, the p-values are not uniformly distributed (KS-test: p<0.01). In fact, 31% of the 189 coefficients have a p-value of less than 0.05. Thus, while the average gender bias is indistinguishable from zero and most judge-specific coefficients are insignificant, some judges seem to favour riders of the same gender while some judges favour riders of the opposite gender. Out of the significant estimates of α_k, roughly half (46%) are positive and half (54%) are negative. This pattern is strikingly similar for male and female judges: 15% of male and 13% of female judges exhibit significant own-gender favouritism ($\chi^2(1)$=0.143, p=0.705) while 17% of male and 16% of female judges exhibit significant opposite-gender favouritism ($\chi^2(1)$=0.067, p=0.800).

[Figure 1 about here]

To assess the impact of gender bias on the riders' final outcomes, it is important to keep two issues in mind: First, as shown in the above analyses, the average null effect of judge gender hides substantial heterogeneity in the size and direction of the judge-specific gender bias. Second, since there are many different nationalities but only two genders, biased judges are more likely to face a rider of their preferred gender than a rider of their preferred nationality. Thus, while the gender bias is, on average, smaller and is exhibited by fewer judges than the nationalistic bias, it will distort a larger number of evaluations for each affected judge.[12] To investigate the quantitative importance of these issues, I adjust the final ranking of riders within each competition, taking the judge-specific gender bias estimated in model (3) into account. That is, for all judges exhibiting significant (p<0.05) judge-specific bias in favour of their own (or the opposite) gender, I deduct their estimated bias from their scores of riders of their own (or the opposite) gender. Computing new rankings based on these

[12] I thank an anonymous reviewer for pointing this out.

"bias-adjusted" scores, I find that the final position changes for 2.8% of all performances and for at least one performance in 14.8% of all competitions. Thus, even though the average gender bias is small and insignificant, and judges are more likely to exhibit significant nationalistic bias than significant gender bias, the gender bias has a non-negligible impact on final outcomes. Repeating the same exercise for the judge-specific nationalistic bias, I find that the final position changes for 5.3% of all performances and for at least one performance in 24.3% of all competitions. This illustrates that, while the average gender bias only amounts to 3% of the average nationalistic bias (0.010 vs. 0.359), the distortion of outcomes generated by the gender biased judges is more than half the size of that generated by the nationalistically biased judges.

Next, I consider the possibility that the effects of gender and nationality may interact, either magnifying or counteracting each other. Thus, in the second column of Table 3, I interact *J&R same nationality* with *J&R same gender*. The interaction term is small and insignificant, indicating that a judge's nationalistic bias does not increase when the rider is of the same gender as the judge. Relatedly, Figure A4 in the Appendix illustrates the relationship between the judge-specific nationalistic bias and the judge-specific gender bias. The correlation between the two types of biases is small and insignificant (ρ=-0.013, p=0.869), indicating that the level of a judge's nationalistic bias is not systematically related to their gender bias.[13,14]

The final column of Table 3 shows that the nationalistic bias is on average 0.101 points, or 34%, larger for female riders than for male riders. Thus, judges seem to favour riders of their own nationality to a greater extent when the rider is female. As shown in Table

[13] Throughout this paper, when I compute correlations involving judge-specific biases, or when I run regressions using the judge-specific bias as dependent variable, I weigh each observation by the inverse of the standard error of the judge-specific bias estimate. In cases like this, when I compute the correlation between two different judge-specific biases, I weigh the observations by the average of the two inversed standard errors. Unless otherwise stated, all results are robust to weighing by the number of observations per judge or using no weights.

[14] The correlation between the judge-specific nationalistic bias and the *absolute size* of the judge-specific gender bias is significant at the 10% level (ρ=0.140, p=0.077). However, this correlation decreases and is rendered insignificant if I weigh each observation by the number of observations per judge or control for the variability of the judge-specific bias estimates.

B2 in the online Appendix, the size of this interaction effect does not differ significantly between female and male judges and it is not driven by observable gender differences in rider or competition characteristics. However, I cannot rule out that this finding is due to gender differences in non-observable characteristics.

4. Heterogeneity

4.1. Type of Competition

Findings from previous studies indicate that the degree of in-group bias is positively correlated with the salience of group identity (e.g. Mullen *et al*., 1992; Eckel and Grossman, 2005; Charness *et al*., 2007; Rand *et al.*, 2009; Chen and Chen, 2011; Shayo and Zussman, 2011). For example, Shayo and Zussman (2011) find that the own-ethnicity bias among judges in Israeli courts increases with the intensity of ethnic conflict in the area surrounding the court. To investigate if the nationalistic bias among dressage judges is associated with the salience of national identity, I make use of the fact that riders compete for their nation only in certain competitions: international championships and team competitions (where a group of riders from the same country compete as a team). In all other competitions the riders represent only themselves. My data include 18 championship competitions (from six different championships) and 65 other competitions that are part of a team event.[15]

The middle section of Figure 2 shows the size of the nationalistic bias estimated by running separate regressions for championships, team events, and other competitions. The nationalistic bias is larger in championships (0.644) and team events (0.501) than in other competitions (0.345).[16] As shown in the first column of Table 4, the difference between championships and other competitions is significant at the 10% level, and the difference between team events and other competitions is significant at the 5% level. The size of the bias

[15] A team event is a major event that includes at least one team competition (see online Appendix D for more information).
[16] In online Appendix C.1 I verify these findings using fixed effects for each unique combination of judge and rider. I show that the positive effect of championships and team competitions is larger for judge-rider pairs of the same nationality than for those of different nationalities.

in championships is substantial, corresponding to 10.7% of the overall standard deviation and 45.4% of the within-performance standard deviation of technical scores in championships. When I deduct this bias from the technical scores from judges of the same nationality as the rider, the final position changes for 11.2% of all championship performances in the data. The final position changes for at least one rider in all championship competitions.

These results are in line with the hypothesis that nationalistic bias increases as the national dimension of the competition becomes more salient. The size of the nationalistic bias does not vary significantly with other observable competition characteristics such as the average final score, the level of the competition or the prize money at stake (see Table B1 in the online Appendix). Thus, the effect of championships and team events does not seem to be driven by differences in monetary stakes or the average quality of riders in the competition.

[Figure 2 and Table 4 about here]

4.2. Number of Rider Nationalities

Identification with national identity might also explain the lack of gender bias among most judges. It is possible that national identity is more salient than gender identity in international competitions, causing nationalistic bias to crowd out any underlying gender bias. If so, there may be more room for gender identity to become salient, and for gender bias to emerge, when few nationalities are represented in a competition. In the second column of Table 4 I interact *J&R same gender* with the number of rider nationalities represented in the competition. The interaction term is negative and significant ($p<0.05$), indicating that the average size of a judge's gender bias decreases by 0.006 points for each additional rider nationality.[17] In the rightmost section of Figure 2, I divide the sample into three groups based on the number of

[17] In additional regressions, I allow for nonlinearities by using other functional forms for the interaction term. When interacting *J&R same gender* with the square root of *No. of rider nationalities*, the coefficient of *J&R same gender* increases to 0.111 ($p=0.039$) and the coefficient of the interaction term is -0.039 ($p=0.043$). Similarly, when using the natural log of *No. of rider nationalities*, the coefficient of *J&R same gender* increases to 0.095 ($p=0.045$) and the coefficient of the interaction term is -0.048 ($p=0.050$).

rider nationalities represented in the competition. Running separate regressions for these groups, I find a statistically significant ($p<0.01$) gender bias of 0.086 in the lowest tercile, i.e. in competitions with only 1-4 rider nationalities. I find no such effect in the other terciles. The third column of Table 4 shows that the gender bias is significantly ($p<0.05$) larger in competitions with 1-4 nationalities than in other competitions.[18] These results are in line with the conjecture that gender bias is, to some extent, crowded out by nationalistic bias in environments with strong nationalistic group identification. As shown in Table B3 in the online Appendix, this effect is not driven by observable differences in competition characteristics such as geographic region, the average quality of riders, the level of the competition or the prize money at stake.

4.3. Judge Characteristics

Are judges from some countries more biased than others? Are older and experienced judges less biased? To answer these questions, I collapse the data on the judge level. As outlined in online Appendix C.2, I find no statistically or economically significant relationship between the judge-specific nationalistic (gender) bias and various measures of nationalism (gender equality) in the judge's home country. Figure A5 in the Appendix illustrates the correlation between the judge-specific bias and the judge's age and experience. Age is not significantly related to either of the judge-specific biases. However, the judge's total number of scores included in the data (a proxy for experience) is negatively correlated with the judge-specific nationalistic bias ($p<0.01$). As a judge's total number of scores increase by one standard deviation (671 scores), the judge-specific nationalistic bias decreases by 11% of a standard deviation (0.07 points).[19] Finally, while the judge-specific nationalistic bias is, on average, larger for female judges (mean=0.52, SD=0.65) than for male judges (mean=0.43, SD=0.56), this difference is not statistically significant ($t=0.88$, $p=0.38$).

[18] Table B4 in the online Appendix shows that using 1-3, 1-5, or 1-6 instead of 1-4 rider nationalities yield similar results.
[19] I obtain this number by regressing the judge-specific nationalistic bias on the judge's age, gender and number of scores.

In online Appendix C.3 I show that the degree of nationalistic bias is not affected by the history of conflicts or bilateral trust between the judge's country and the countries of the riders who compete *against* a rider from the judge's country.

5. Behavioural Interpretation of Nationalistic Bias

The previous sections show that judges systematically give higher scores to riders of their own nationality. This effect is non-negligible in size and impacts the final ranking of riders in many competitions. However, the relevance of these findings for other contexts hinges on the mechanisms behind the judges' behaviour. Does the observed effect primarily reflect unconscious nationalistic bias? Or can part of the effect be explained by strategic behaviour, personal connections between the judges and riders, preferences for certain riding styles, or omitted variables? In this section, I empirically address these questions.

5.1. Strategic Voting

Several features of dressage support the notion of biases being primarily unconscious. First, judges face a relatively high degree of monitoring, as competitions are observed by audiences and often televised, and all scores are saved in a public database. Second, the appointment and promotion of judges is not managed by the judges' respective national federations. Instead, the International Federation for Equestrian Sports governs these processes with the aim of ensuring consistent and unbiased scoring. Third, the scoring procedure is quite fast-paced, leaving less room for deliberate and reflective choices. A performance lasts less than six minutes, during which the judge must provide 37 scores (the final score is the mean of these).

Moreover, if judges consciously aim to maximize the outcomes of their countrymen, I expect this effect to increase when the judge has a realistic chance of improving the rider's final position. Rows (1) - (5) of Table A1 in the Appendix show that the size of the nationalistic bias is not significantly affected by how close a rider is to winning or gaining/losing a position in the competition. Rows (6) - (7) verify that this finding holds when

restricting the sample to European Grand Prix Freestyle competitions, in which the final position is particularly important as it contributes to the World Cup ranking. In rows (8) - (10) I turn to competitions that also act as qualifiers, meaning that the riders who rank 1-15 are eligible to participate in a subsequent competition within the same event.[20] If judges vote strategically, I expect the nationalistic bias to increase when a rider is close to making the cut-off. I find no such effect.

Finally, I expect judges who strategically favour a rider to be biased *against* that rider's closest competitors. Thus, I restrict the sample to scores where the judge and the rider are of different nationalities. Then I run regression model (1), adding an explanatory variable indicating if at least one of the rider's closest competitors (those within one [two] positions of the rider) is of the same nationality as the judge. The coefficient of this variable is small (0.009 [0.025]) and insignificant.

In sum, these analyses provide no indication of judges consciously and strategically favouring riders of their own nationality.

5.2. Personal Connections between Judges and Riders

As outlined in online Appendix D, the rules of dressage prevent judges from having any conflict of interest concerning the competitors. For example, judges are not allowed to train a rider, own a horse, or have a close personal relationship with a rider in the competition. In the final rows of Table A1, I explore if other forms of personal connections between judges and riders may matter. First, I assume that riders and judges of the same nationality are more likely to know each other well if the country's dressage community is small. Thus, if part of the observed bias is driven by personal connections, the nationalistic bias should be stronger in countries with few judges and riders. Rows (11) - (12) show that this is not the case.

[20] In my data, there is no information on whether a competition was a qualifier. Thus, in this analysis I restrict the sample to the 75 competitions I can clearly identify as qualifiers. The sample is thus reduced to 7141 scores provided by 85 judges to 489 riders for 1443 performances.

Similarly, connections between judges and riders of the same nationality might be stronger when the country is ethnically, religiously and linguistically homogenous. Thus, in rows (13) – (15) I interact *J&R same nationality* with the country's scores on the fractionalization indices developed by Alesina *et al.* (2003). The interaction effects are small and insignificant.

Finally, I assume that the judge is more likely to have a personal connection with the rider or other stakeholders in the horse if the judge and the horse are of the same nationality.[21] Thus, I run regression model (1) again, adding an explanatory variable indicating if the judge and the horse are of the same nationality (*J&H same nationality*) and an interaction between this variable and *J&R same nationality*. The coefficient of *J&R same nationality* is 0.377 (SE=0.057, p<0.01), the coefficient of *J&H same nationality* is 0.161 (SE=0.047, p<0.01) and the coefficient of the interaction term is -0.161 (SE=0.085, p=0.06). This indicates that when the judge and the rider are of different nationalities, judges give on average 0.161 more points to horses of their own nationality. However, there is no such effect when the judge and the rider are of the same nationality. Thus, the positive effect of *J&R same nationality* is not influenced by the horse's nationality.

Taken together, these tests provide no evidence that the nationalistic bias is driven by personal connections between judges and riders.

5.3. Nation-Specific Preferences for Riding Styles

Do judges give higher scores to riders of their own nationality because they share a preference for a particular style of riding? Several features of the sport speak against this conjecture. First, judges undergo extensive and standardized international training, thereby learning to conform to a common standard. Second, dressage is a highly international sport. International judges, riders and trainers travel all over the world to judge, compete and train, reducing the scope for

[21] The horse's nationality indicates the country that the horse was bred in and/or that the horse is, or has been, owned by an individual from that country. I retrieved the horse's nationality from the horse's international ID number.

nation-specific preferences. Third, nation-specific preferences for certain riding styles cannot explain why the nationalistic bias increases in championships and team events.

To further assess the importance of nation-specific preferences, I turn to Grand Prix Freestyle (GPF) competitions, in which riders are given both technical and artistic scores. The guidelines for how to judge the artistic quality of a performance are substantially less formalized than for the technical scores, providing more room for judges to express their individual preferences. Thus, if nation-specific preferences for riding styles play a role, I expect to find a larger bias in artistic scores. However, restricting the analysis to GPF competitions, the coefficient of *J&R same nationality* is slightly smaller when using the artistic score as outcome variable (β=0.326, SE=0.067) than when using the technical score (β =0.409, SE=0.046). This difference is not statistically significant (Z=1.021, p=0.307).

5.4. Omitted Variables

All regressions in this paper include judge fixed effects, thereby controlling for average differences in leniency between judges. Thus, if judges from certain countries are, on average, more generous or harsh in their scoring, this will not influence the results. However, what if judges of different nationalities differ systematically with respect to other characteristics, such as age and experience, causing them to discriminate between riders of different nationalities? To explore this, I add controls for rider nationality interacted with the judge's age, gender and experience (the number of performances in the data that are scored by the judge) to model (1). The resulting change in the coefficient of *J&R same nationality* (from 0.359 to 0.354) is small and insignificant (Z=0.121, p=0.904). Thus, I find no evidence that my findings are driven by differences in (observable) judge characteristics other than nationality.

6. Indirect Bias

Gender and nationality represent relatively permanent groups in the sense that individuals rarely change their gender or nationality. In contrast, the judging panel of a dressage

competition only exists for a few days.[22] In this section, I explore if judges are influenced by the characteristics of their fellow judges on the panel. In particular, I investigate if judges favour athletes that are of the same nationality as some of the other panel members.

6.1. Estimation

First, I need to empirically distinguish between:

(i) Direct nationalistic bias (β_{direct}): Judges favouring riders of their own nationality, and

(ii) Indirect nationalistic bias ($\beta_{indirect}$): Judges favouring riders of the same nationality as another judge represented on the judging panel.

In model (1), the coefficient of *J&R same nationality* measured the difference between the average score from judges of the same nationality as the rider, and the average score from judges of another nationality. Thus, this coefficient captured the difference between the direct and the indirect biases: β=β_{direct}−$\beta_{indirect}$. To distinguish between β_{direct} and $\beta_{indirect}$, I estimate a modified version of a model introduced by Zitzewitz (2006): [23]

$$s_{jrcp} = \alpha \cdot J\&R\ same\ gender + \beta_{direct} \cdot J\&R\ same\ nationality \quad (4)$$

$$+ \beta_{indirect} \cdot R\ same\ nationality\ as\ other\ J + \gamma_j + \lambda_r + \mu_c + e_{jrcp}.$$

The model includes fixed effects for judges (γ_j), riders (λ_r) and competitions (μ_c), and the standard errors are clustered on three levels: judge, rider and performance. The indicator variable *R same nationality as other J* takes the value 1 if the judge and the rider are of different nationalities but the rider is of the same nationality as some other judge on the panel. To estimate this model, I must assume that there is no correlation between the composition of

[22] The judging panel is often the same or similar throughout an event consisting of several consecutive one-day competitions.
[23] Zitzewitz (2006) explores how judges in ski jumping and figure skating are affected by the nationality of the other judges on the panel. He uses the term "compensating bias" to describe what I call "indirect bias".

judge nationalities represented on the panel and the quality of a rider's performance, except for what is captured by the fixed effects.[24]

6.2. Main Results

In the first two columns of Table 5, I estimate model (4) with and without the explanatory variable *R same nationality as other J*. When included in the regression, the coefficient of this variable is 0.277 and statistically significant ($p<0.01$), indicating that judges systematically reinforce each other's biases by giving higher scores to riders of the same nationality as the other judges on the panel. Moreover, the coefficient of *J&R same nationality* increases significantly, from 0.416 to 0.623 ($Z=2.132$, $p=0.033$), when allowing for indirect bias. The reason for this increase is that the coefficient in the first column compares the score from judges of the same nationality as the rider to the score from judges of other nationalities. However, the positive indirect bias implies that, on average, *all* judges on the panel are biased in favour of riders with at least one judge of the same nationality. Thus, in the first column, the scores from the comparison group of judges (those of a different nationality than the rider) are upward biased.[25]

When allowing for indirect bias, the size of the nationalistic bias in Table 5 increases from 8.2% to 12.4% of one standard deviation of technical scores. In terms of the within-performance standard deviation, the bias increases from 27.5% to 41.3%. Moreover, the total effect on a rider's final score of having at least one judge of the same nationality on the panel increases considerably. To illustrate this, consider the average increase in a rider's final score when going from zero to one judge of the same nationality. Assuming no indirect bias, this

[24] As noted by Zitzewitz (2006), this identifying assumption is violated if riders perform better, and are more likely to face a judge of their own nationality, in competitions located in their home country. My estimate of indirect bias is robust to excluding all riders competing in their home country.

[25] Note that the coefficient of *J&R same nationality* captures β in column (1) and β_{direct} in column (2). As explained in Section 6.1, $\beta=\beta_{direct}-\beta_{indirect}$, or, equivalently, $\beta_{direct}=\beta+\beta_{indirect}$. Since $\beta_{indirect}>0$ it follows that $\beta_{direct}>\beta$.

increase is 0.083 points.[26] Taking the indirect bias into account, the equivalent increase is 0.346 points, i.e. almost five times as large[27,28]

In online Appendices C.4 and C.5, I show judge-specific indirect biases and heterogeneity analyses. Among other things I find that, like the direct bias, the indirect bias increases in championships and team competitions.

[Table 5 about here]

6.3. Behavioural Interpretation of Indirect Bias

The indirect bias may reflect unconscious bias in favour of the temporary in-group generated by the judging panel. Alternatively, it could reflect conscious, strategic behaviour. Specifically, judges might engage in vote trading, try to get on good terms with certain influential judges, or be averse to being outliers. In this section I address these possibilities.

Vote trading may occur either *within competitions* (judges on the same panel favour riders of each other's nationalities in the competition) or *across competitions* (judge A favours a rider of judge B's nationality in one competition, and Judge B reciprocates in another competition). If vote trading within competitions accounts for the indirect bias, it should only affect the behaviour of judges who have at least one rider of their own nationality in the competition. Thus, I interact *R same nationality as other J* with the variable *J Same Nationality as other R*. This variable indicates if the judge and the rider are of different nationalities, but the judge is of the same nationality as some other rider in the competition. As shown in column (3) of Table 5, the indirect bias remains large and significant for judges who do not have a rider of their own nationality in the competition. Hence, vote trading within competitions does not seem to account for the indirect bias.[29]

[26] (0.416+4*0)/5=0.083
[27] (0.623+4*0.277)/5=0.346
[28] To estimate the impact on riders' final positions, I deduct 0.623 from all scores for which *J&R same nationality=1*, and 0.277 from all scores for which *R same nationality as other J=1*. As a result, the final position in the competition changes for 16.3% of all performances, and at least one position changes in 45.6% of all competitions.
[29] Note that I cannot rule out the existence of vote trading within competitions. The small and insignificant interaction term may capture two opposing effects: First, vote trading within competitions predicts a *positive* interaction effect. Second, when

In theory, vote trading across competitions should not account for the indirect bias. Since the scores of all judges are public information, two judges can trade votes across competitions without ever being on the same panel. However, in practice, being on the same panel may facilitate vote trading across competitions. If vote trading across competitions drives the indirect bias, the effect should be restricted to pairs of judges who participate in the same panel more than once within a limited time period. Thus, in Table A2 in the Appendix, I interact *R same nationality as other J* with different measures of how often the judge providing the score is on the same panel as the judge who is of the same nationality as the rider. Results show that the indirect bias does increase when the judges have been on many panels together and when their most recent or upcoming meeting takes place close in time to the current competition. However, the coefficient of *R same nationality as other J* remains large and significant, indicating that the effect is not restricted to pairs of judges who meet repeatedly. Hence, while I cannot rule out that vote trading may contribute to the indirect bias, my results suggest that it does not solely account for it.

Relatedly, judges may strategically favour riders of the same nationality as the most important judge(s) on the panel. For instance, they may seek approval from other judges that are influential within the sport or likely to judge certain important competitions in the future. In column (4) of Table 5 I explore this possibility by interacting *R same nationality as other J* with the following characteristics of the judge on the panel who is of the same nationality as the rider: (i) the total number of competitions that he/she has served on, (ii) whether he/she has ever judged a championship or world cup final, and (iii) whether he/she has reached the highest possible judge ranking. All three coefficients are statistically insignificant, but the coefficient of the indirect bias decreases by 40% to 0.165. While this indicates that the

there is a rider of their own nationality in the competition, judges may try to improve the outcome of that rider by punishing riders of the same nationality as other judges, generating a *negative* interaction effect. Thus, while I rule out that vote trading within competitions *drives* the observed indirect bias, I cannot rule out that vote trading within competitions *exists*.

indirect bias may be partly related to the importance of the other judges, the decrease is not statistically significant (Z=0.904, p=0.366).

Finally, as pointed out by Zitzewitz (2006), being an outlier may harm a judge's reputation. Thus, judges may seek to conform to the expected average score from the other judges on the panel. If judges expect the others to be nationalistically biased, a desire to conform may cause indirect bias. If so, I expect judges to increase their scores more when they expect the judge who is of the same nationality as the rider to exhibit a large nationalistic bias. Thus, I interact *R same nationality as other J* with the judge-specific nationalistic bias (as estimated by equation (2)) of the other judge on the panel who is of the same nationality as the rider. Assuming that judges have a fairly accurate perception of the size of other judges' biases, a conformity effect would imply a positive interaction. However, as shown in column (5) of Table 5, the coefficient of the interaction term is negative and insignificant.

Taken together, these results suggest that the indirect bias is not primarily driven by conscious, strategic behaviour in the form of vote trading, seeking approval from influential colleagues, or the fear of being an outlier. A remaining explanation that cannot be ruled out is that the judging panel creates a temporary group identity, and that the indirect bias reflects unconscious in-group bias in favour of this temporary in-group.

7. Conclusions

This paper investigates in-group biases among judges in international dressage competitions. On average, while judges are not biased in favour of either gender, they exhibit a sizeable bias in favour of athletes of their own nationality. The nationalistic bias is particularly striking given that these judges are trained, experienced and constantly monitored professionals. Most likely, a bias needs to be strong and resilient to penetrate this expert group of evaluators. Nevertheless, the variation across types of competitions suggests that the nationalistic bias is also malleable and responds to situational factors.

While most existing studies on in-group bias focus on one single bias, this study contributes by investigating both gender bias and nationalistic bias in the same setting. Incorporating multiple group identities in the empirical research on in-group bias is important since in real life individuals belong to multiple groups simultaneously. Nationality seems to play a larger role than gender in influencing the behaviour of international dressage judges. My analysis suggests that this might partly be due to nationalistic bias crowding out gender bias in international settings. An important avenue for future research is to further study the interplay between different in-group biases in environments where several identities compete.

This study also speaks to the economics of discrimination. However, one limitation of my data is the lack of an objective measure of performance quality. This prevents me from distinguishing between positive and negative discrimination, i.e. whether judges are primarily biased in favour of riders of their own nationality or against riders of other nationalities. This also makes it difficult to identify statistical discrimination, especially if all judges hold similar beliefs about average quality differences between groups of riders. I compare the scores from the in-group members of the panel to the scores from the out-group members of the panel for each performance. Thus, if all judges discriminate against the same group of riders (e.g. if both male and female judges discriminate against women), this would not be possible to detect in the data. My estimates of in-group bias indicate either taste-based discrimination or biased beliefs about rider ability that systematically vary across different groups of judges. Since all judges observe the rider's full performance, and thus are perfectly informed about performance quality, I find it unlikely that biased beliefs about ability play a major role.

The indirect bias implies that going from zero to one in-group member on a committee can have a large impact on outcomes, as it affects the verdicts of all committee

members.[30] Thus, there is a risk of severely underestimating the total effect of group membership on outcomes if only taking the bias of the in-group members into account. Moreover, while dressage judges engage in simultaneous voting without prior discussions, many decisions by evaluation committees are reached through joint consensus following a group discussion. In such committees, the scope for indirect bias is likely to be even larger, since the judges can argue for their case and the social cost of disagreeing may increase.

This study adds to a growing literature using the richness and availability of sports data to explore labour market related phenomena (e.g. Duggan and Levitt, 2002; Pettersson-Lidbom and Priks, 2010; Price and Wolfers, 2010; Parsons *et al.*, 2011). Among previous studies using sports data, the current analysis most closely resembles Zitzewitz's (2006) interesting investigation of direct and indirect nationalistic bias among judges in Olympic winter sports. Some of my results replicate his findings. Judges in both figure skating and ski jumping, like dressage judges, favour competitors of their own nationality. In ski jumping, the nationalistic bias is also largest in championships and team competitions. Moreover, Zitzewitz (2006) finds a positive and significant indirect nationalistic bias among judges in figure skating, which he interprets as a sign of vote trading. In contrast to this interpretation, and based on the analysis presented in Section 6.2, I argue that the indirect bias among dressage judges is not primarily driven by vote trading.

These results are potentially relevant to policy makers, managers, and organisational designers who are interested in reducing in-group favouritism in evaluation situations. Many labour market evaluations are similar to dressage competitions in the sense that they take place in a competitive and international mixed-gender setting and involve high stakes. Examples of such evaluations include international recruitment or promotion committees, courts, universities and legislative bodies. The robust evidence of substantial nationalistic bias

[30] Similarly, studying the racial composition of jury pools in the US, Anwar *et al*. (2012) find that going from zero to one black member of the jury pool changes the behaviour of all jury members. In particular, including one black member completely eliminates the racial gap in conviction rates.

among dressage judges – an expert group of evaluators facing unusually high levels of monitoring and accountability – indicates that similar processes are likely to operate in most other evaluation settings in the labour market.

One obvious way of reducing in-group bias in evaluation decisions is to balance panels or exclude evaluators of the same nationality as competitors/applicants. For instance, soccer referees are prohibited from refereeing matches involving their own country. However, this type of reform would be difficult to implement in dressage (and similar settings) due to the large number of nationalities participating. Another way to reduce the impact of nationalistic judging, used in many other sports, is truncation of extreme scores. However, truncation would not address the positive indirect bias among the remaining judges. Similarly, due to the indirect bias, it would not suffice to remove the score from the in-group judge or correct their score downward to adjust for the anticipated bias. One potentially promising avenue for reducing in-group bias in evaluation committees may be to decrease the salience of the undesirable group identity (e.g. nationality) by emphasizing and increasing the salience of another, less harmful, group identity (e.g. corporate identity or global citizenship).

Table 1
Sample Size

	All	Men	Women
No. of riders	1,215	384	831
No. of rider nationalities	61	47	54
No. of judges	191	86	105
No. of judge nationalities	42	29	35
No. of competitions	1,533	1,344	1,510
No. of championships	18	18	18
No. of performances	18,201	6,113	12,088
No. of performances in championships	626	246	380
No. of scores	90,626	30,470	60,156
No. of scores in championships	3,348	1,318	2,030
Average no. of riders in a competition	16.84	17.86	16.32
Average no. of judges in a competition	4.98	4.98	4.98
Share of scores with judge of same nationality as the rider	17.4%	17.3%	17.4%
Share of scores with judge of same gender as the rider	48.7%	57.2%	44.4%

Notes. The gender-specific statistics of *No. of judges* and *No. of judge nationalities* refer to the gender of the judge. All other gender-specific statistics refer to the gender of the rider.

Table 2

Descriptive Statistics

	RIDERS				JUDGES	
	All	Male	Female	Championships	Male	Female
Technical score						
Mean	65.23	65.46	65.11	68.79	65.38	65.04
(overall SD)	(5.02)	(4.81)	(5.12)	(6.01)	(5.03)	(5.00)
[within performance SD]	[1.51]	[1.52]	[1.51]	[1.42]	-	-
Min - Max	22.9 - 95.8	27.7 - 95.0	22.9 - 95.8	22.9 – 92.0	22.9 - 95.8	25.5 - 93.0
p25 - p75	62.2 - 68.2	62.6-68.3	62.0-68.1	65.1 - 72.0	62.3 - 68.3	62.0 - 68.0
Artistic score						
Mean	71.38	71.84	71.16	80.68	71.63	71.06
(overall SD)	(6.53)	(6.31)	(6.61)	(6.21)	(6.52)	(6.52)
[within performance SD]	[2.20]	[2.21]	[2.20]	[2.21]	-	-
Min - Max	28.5 - 97.0	28.5 - 97.0	31.5 - 96.0	58.0 - 97.0	28.5 - 97.0	32.5 - 96.0
p25 - p75	67.0-75.0	68.0 - 75.0	67.0 - 75.0	76.0 - 85.0	68.0 - 75.0	67.0 - 75.0
Age						
Mean	39.10	40.49	38.39	37.41	58.79	58.70
(overall SD)	(10.06)	(9.99)	(10.02)	(9.04)	(6.85)	(6.77)
Min - Max	15 - 74	18 - 72	15 - 74	17 - 69	36 - 72	29 - 71
p25 - p75	31 - 46	33 - 48	30 - 46	30 - 44	55 - 63	55 - 64
Nationality						
Western European	62%	72%	56%	78%	78%	58%
Central European	12%	6%	15%	10%	15%	7%
North American	15%	10%	17%	6%	5%	19%

Notes. p25 refers to the 25th percentile and p75 refers to the 75th percentile.

Table 3
Baseline Results

	(1)	(2)	(3)
J&R same nationality	0.359***	0.348***	0.293***
	(0.026)	(0.034)	(0.040)
J&R same gender	0.010	0.006	0.008
	(0.017)	(0.019)	(0.017)
J&R same nationality X J&R same gender		0.022	
		(0.041)	
J&R same nationality X Female rider			0.101**
			(0.042)
N	90,626	90,626	90,626
Mean dependent variable	65.229	65.229	65.229
Std. of dependent variable	5.023	5.023	5.023

Notes. All estimates are based on OLS regressions with fixed effects for performance and judge. The dependent variable is the technical score provided by a judge for a performance. Standard errors clustered on judge and rider are shown in parentheses. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table 4
Heterogeneity

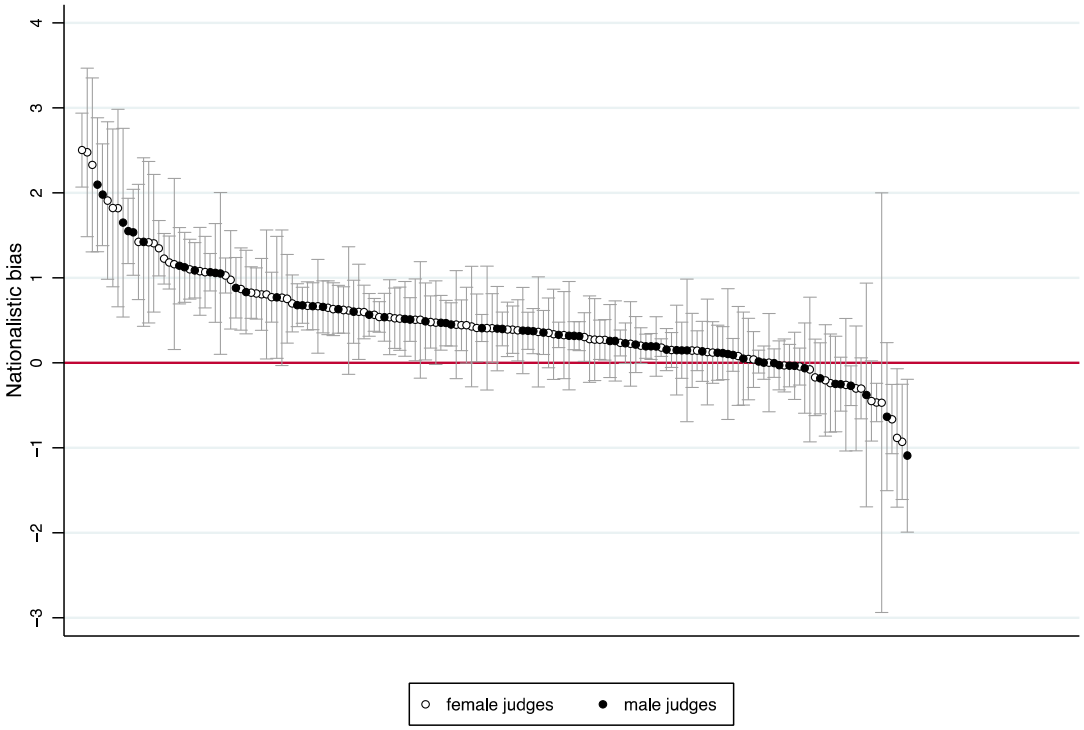
	(1)	(2)	(3)
J&R same nationality	0.345***	0.358***	0.358***
	(0.027)	(0.026)	(0.027)
J&R same gender	0.010	0.056*	0.082**
	(0.017)	(0.030)	(0.034)
J&R same nationality *interacted with:*			
Championship	0.252*		
	(0.150)		
Team event	0.167**		
	(0.082)		
J&R same gender *interacted with:*			
No. of rider nationalities		-0.006**	
		(0.003)	
5-8 rider nationalities			-0.096**
			(0.041)
9-32 rider nationalities			-0.102**
			(0.044)
N	90,626	90,626	90,626
Mean dependent variable	65.229	65.229	65.229

Notes. All estimates are based on OLS regressions with fixed effects for performance and judge. The dependent variable is the technical score provided by a judge for a performance. Standard errors clustered on judge and rider are shown in parentheses. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

Table 5
Indirect Bias

	(1)	(2)	(3)	(4)	(5)
J&R same nat. (β_{direct})	0.416***	0.623***	0.627***	0.636***	0.625***
	(0.041)	(0.088)	(0.086)	(0.087)	(0.088)
R same nat. as other J ($\beta_{indirect}$)		0.277***	0.268***	0.165*	0.301***
		(0.083)	(0.083)	(0.092)	(0.089)
R same nat. as Other J *interacted with:*					
J same nat. as other R			0.021		
			(0.057)		
Other J: No. of competitions				-0.000	
				(0.009)	
Other J: Important competitions				0.032	
				(0.134)	
Other J: High status				0.210	
				(0.130)	
Other J: Nationalistic bias					-0.063
					(0.104)
J same nat. as other R			0.007		
			(0.025)		
N	90,626	90,626	90,626	90,626	90,547
Mean dependent variable	65.228	65.228	65.228	65.228	65.234

Notes. All estimates are based on regressions including fixed effects for judge, rider and competition, and controls for *J&R same gender*. The dependent variable is the technical score provided by a judge for a performance. Standard errors clustered on performance, rider and judge are shown in parentheses. *R same nat. as other* J=1 if the judge and the rider are of different nationalities but some other judge(s) on the panel is of the same nationality as the rider. *J same nat. as other R*=1 if the judge and the rider are of different nationalities but the judge is of the same nationality as some other rider(s) in the competition. *Other J: No. of competitions* indicates the total number of competitions (in the sample) scored by the other judge on the panel who is of the same nationality as the rider. If several judges are of the same nationality as the rider, I use the maximum value. *Other J: Important competitions*=1 if at least one other judge on the panel, who is of the same nationality as the rider, has judged a championship or a world cup final (in the sample). *Other J: High status*=1 if at least one other judge on the panel, who is of the same nationality as the rider, has achieved an FEI ranking of five stars. *Other J: Nationalistic bias* indicates the degree of nationalistic bias of the other judge on the panel who is of the same nationality as the rider, estimated using model (2). * $p<0.1$; ** $p<0.05$; *** $p<0.01$.



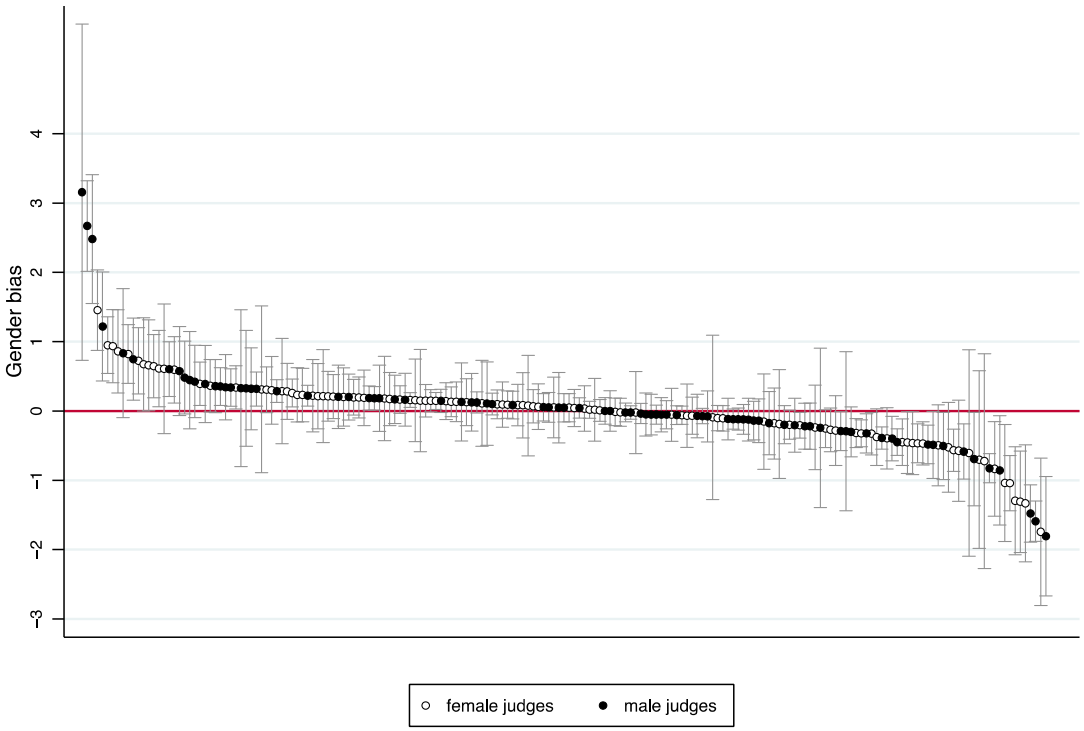


Fig. 1. *Judge-Specific Nationalistic Bias and Gender Bias*

Notes. Each point represents a judge-specific estimate of the degree of nationalistic (gender) bias, i.e. the coefficient $\beta_k(\alpha_k)$ from model 2(3). 95% confidence intervals.

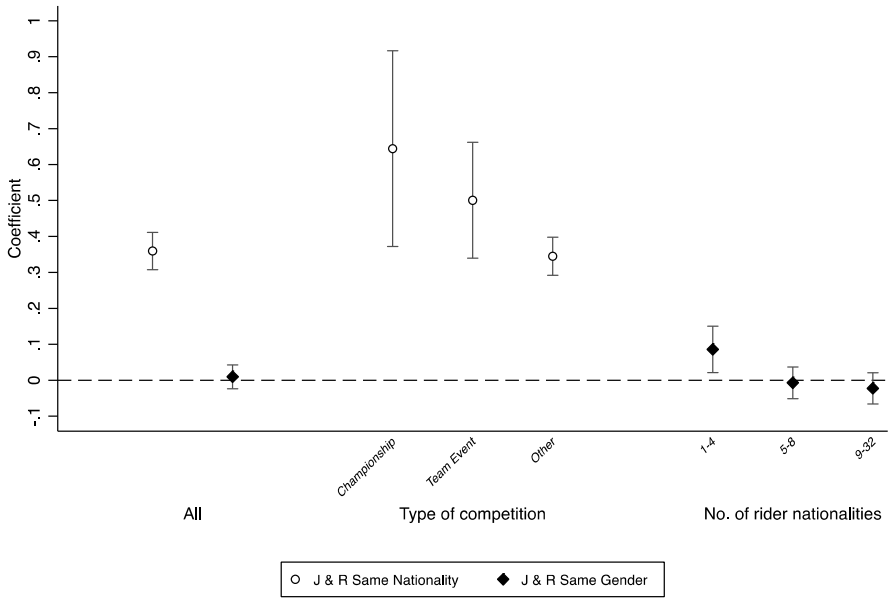


Fig. 2. *Gender Bias and Nationalistic Bias across Types of Competitions*

Notes. The circles represent the degree of nationalistic bias (the coefficient β from model (1)) and the diamonds represent the degree of gender bias (the coefficient α from model (1)). The lines indicate 95% confidence intervals. The coefficients for different subsamples (types of competitions and no. of rider nationalities) are estimated running separate regressions for each subsample.

Appendix A. Additional Tables and Figures

Table A1

Behavioural Interpretation of Nationalistic Bias

		Coefficients			
	Variable	J&R same nat.	J&R same nat. X '*Variable'*	Sample	N
(1)	Final position in the competition	0.354*** (0.041)	0.001 (0.004)	All	90,538
(2)	Distance from winner	0.403*** (0.048)	-0.006 (0.006)	All	82,883
(3)	I [Distance from winner<1]	0.364*** (0.027)	-0.062 (0.081)	All	82,883
(4)	Distance to closest competitors	0.372*** (0.029)	0.000 (0.010)	All	75,618
(5)	I[Distance to closest competitors<1]	0.372*** (0.035)	-0.003 (0.040)	All	75,618
(6)	Distance to closest competitors	0.416*** (0.073)	0.017 (0.026)	European Freestyle	13,528
(7)	I[Distance to closest competitors<1]	0.462*** (0.066)	0.070 (0.112)	European Freestyle	13,528
(8)	Points from cut-off	0.263** (0.110)	0.011 (0.020)	Qualifying GP	7,076
(9)	I[rank 15-16]	0.343*** (0.057)	-0.250 (0.214)	Qualifying GP	7,141
(10)	I[rank 14-17]	0.355*** (0.061)	-0.199 (0.148)	Qualifying GP	7,141
(11)	Number of dressage riders in National Federation /1000	0.358*** (0.038)	0.004 (0.134)	All	90,626
(12)	Number of dressage judges in National Federation /10	0.390*** (0.065)	-0.024 (0.047)	All	90,626
(13)	Country's ethnic fractionalization	0.341*** (0.042)	0.078 (0.158)	All	90,626
(14)	Country's religious fractionalization	0.418*** (0.086)	-0.099 (0.147)	All	90,626
(15)	Country's linguistic fractionalization	0.325*** (0.047)	0.121 (0.155)	All	90,626

Notes. Each row represents a separate OLS regression with fixed effects for performance and judge and controlling for *J&R same gender*. The dependent variable is the technical score provided by a judge for a performance. The explanatory variables of interest are *J&R same nationality* and an interaction between *J&R same nationality* and the variable indicated in the first column. Standard errors clustered on judge and rider are shown in parentheses. *Distance from winner* = Final score of the competition winner – Final score of the rider. *Distance to closest competitors* = Final score of the rider positioned immediately ahead – Final score of the rider positioned immediately behind. *Points from cut-off* indicates how close the rider was to another outcome (i.e., for those who made the cut-off it is the distance to the rider who ranks 16th, and for those who did not make the cut-off it is the distance to the rider who ranks 15th). The variables indicating the number of riders and judges in the National Federation are obtained from https://data.fei.org. This number indicates the total number of riders (judges) of each nationality registered in the FEI database in 2012 (2016). The fractionalization indices are from Alesina *et al.* (2003).

Table A2

Indirect Bias: Effects of Judges' Meetings in Previous/Subsequent Competitions (Excluding the First and Last Six Months of the Sample)

	(1)	(2)	(3)	(4)	(5)
J&R same nat. (β_{direct})	0.648***	0.655***	0.655***	0.651***	0.652***
	(0.084)	(0.083)	(0.084)	(0.084)	(0.084)
R same nat. as other J ($\beta_{indirect}$)	0.318***	0.213***	0.179**	0.205***	0.175**
	(0.076)	(0.082)	(0.082)	(0.078)	(0.082)
R same nat. as other J *interacted with:*					
Judge's no. of meetings with other J		0.006***			
		(0.002)			
Most recent meeting: same day			0.156**		
			(0.066)		
Most recent meeting: within 1-7 days			0.061		
			(0.091)		
Most recent meeting: within 8-183 days			0.031		
			(0.070)		
No. of meetings within 3 months = 1				0.058	
				(0.050)	
No. of meetings within 3 months = 2				0.001	
				(0.066)	
No. of meetings within 3 months: ≥ 3				0.212***	
				(0.074)	
No. of meetings within 6 months: = 1					0.058
					(0.058)
No. of meetings within 6 months: = 2					0.018
					(0.075)
No. of meetings within 6 months: ≥ 3					0.173**
					(0.070)
N	68,888	68,888	68,888	68,888	68,888
Mean dependent variable	65.192	65.192	65.192	65.192	65.192

Notes. All estimates are based on regressions including fixed effects for judge, rider and competition, and controls for *J&R same gender*. The dependent variable is the technical score provided by a judge for a performance. Standard errors clustered on performance, rider and judge are shown in parentheses. *R same nat. as other J* = 1 if the judge and the rider are of different nationalities but some other judge(s) on the panel is of the same nationality as the rider. *Judge's no. of meeting with other J* indicates the total number of other competitions in the sample in which the judge is on the same panel as at least one of the judges who are of the same nationality as the rider. *Most recent meeting: within X* = 1 if the most recent or subsequent competition in which the judge is on the same panel as at least one of the judges who are of the same nationality as the rider takes place within X days of the current competition. *No. of meetings within Y months: X* indicates the number of other competitions in the sample, taking place within Y months of the current competition, in which the judge is on the same panel as at least one of the judges who are of the same nationality as the rider. * $p<0.1$; ** $p<0.05$; *** $p<0.01$.

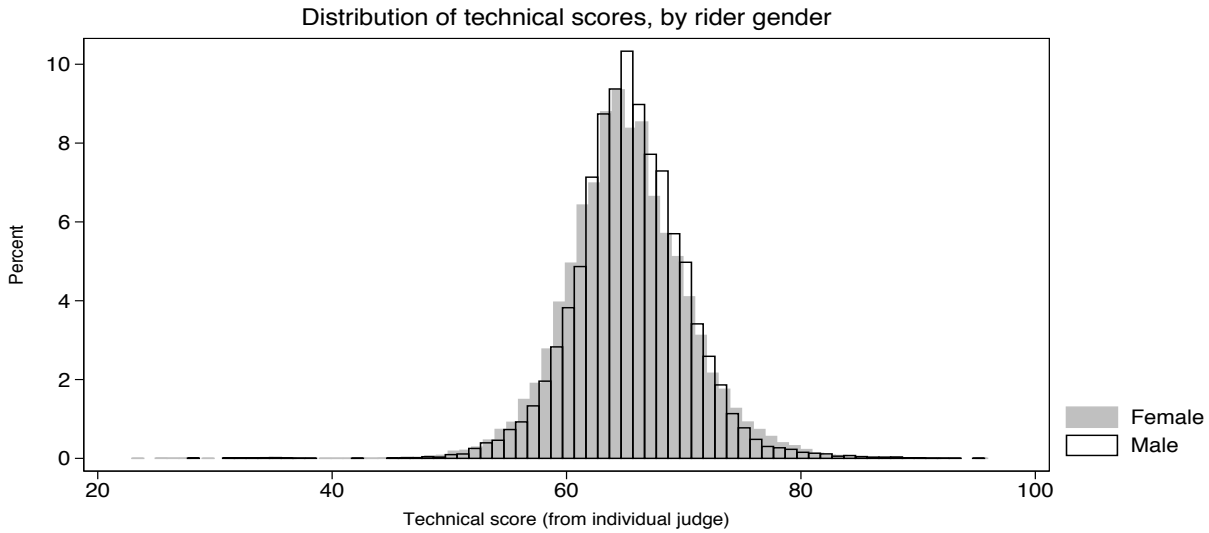


Fig. A1. *Distribution of Technical Scores by Rider Gender*

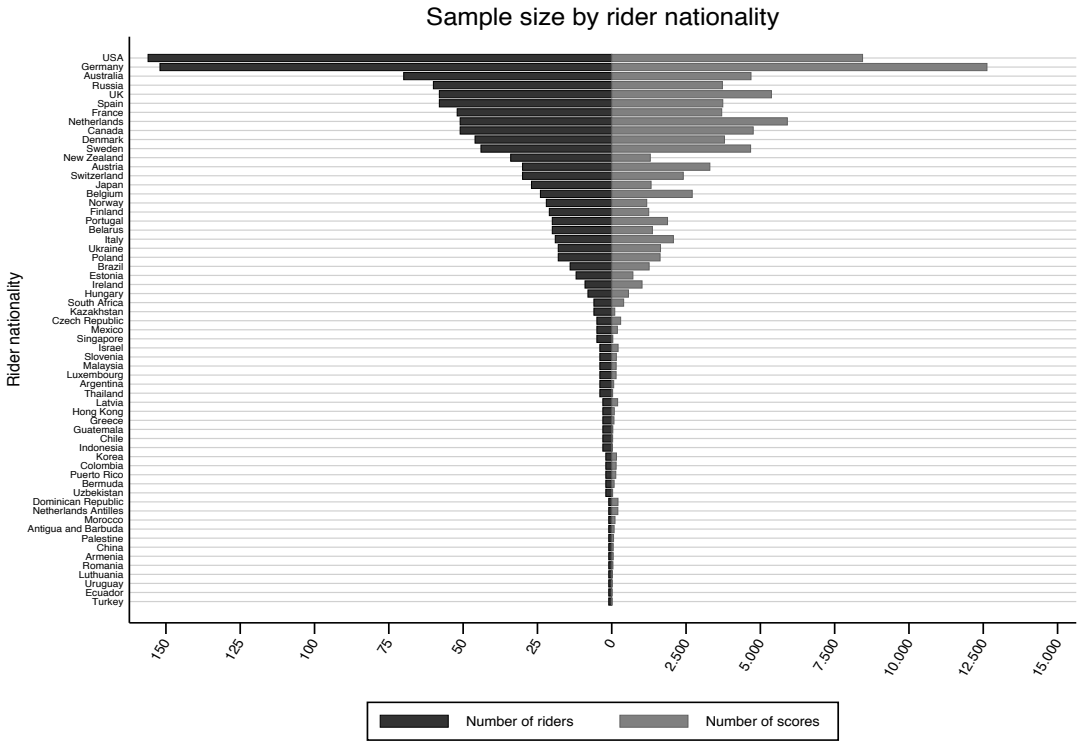


Fig. A2. *Sample Size by Rider Nationality*

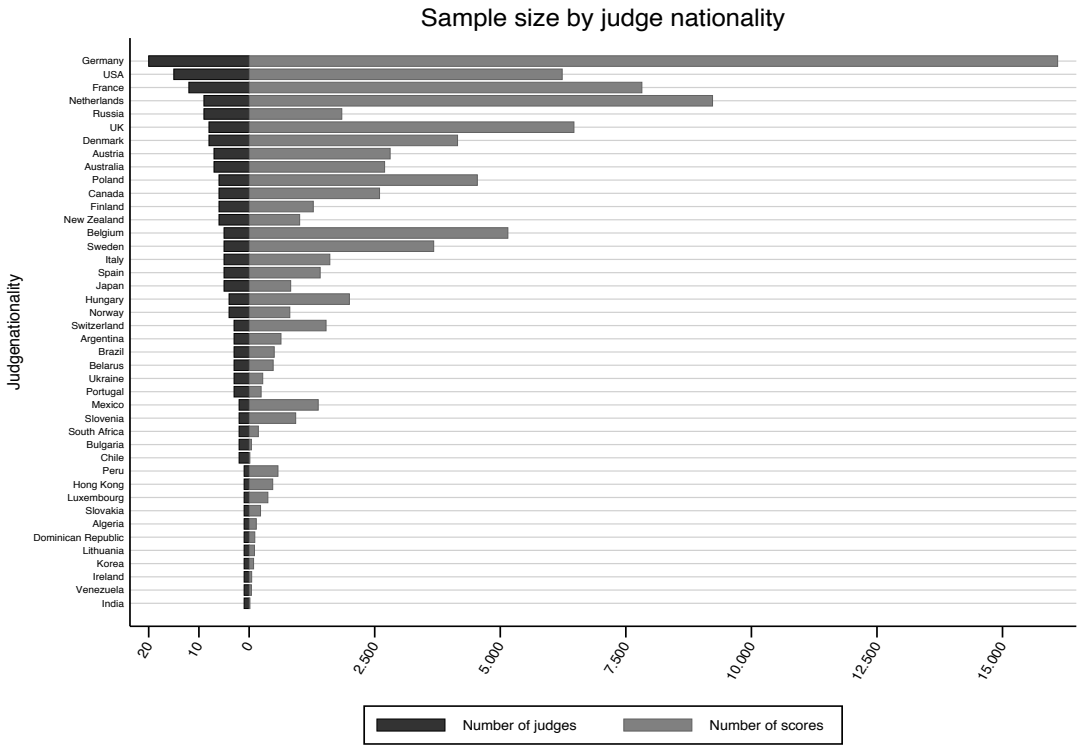


Fig. A3. *Sample Size by Judge Nationality*

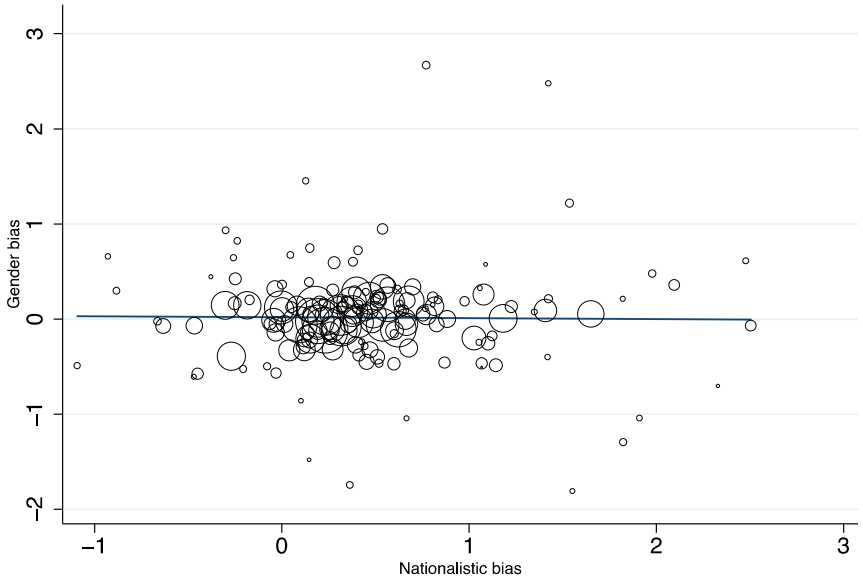


Fig. A4**.** *Correlation between Judge-Specific Gender Bias and Judge-Specific Nationalistic Bias*

Notes. Each circle represents a judge-specific estimate of gender bias (α_k from model (3)) and nationalistic bias (β_k from model (2)). The size of the circles is proportional to the total number of observations in the data for each judge. The line shows the predicted value of the gender bias from a linear regression of the gender bias on the nationalistic bias.

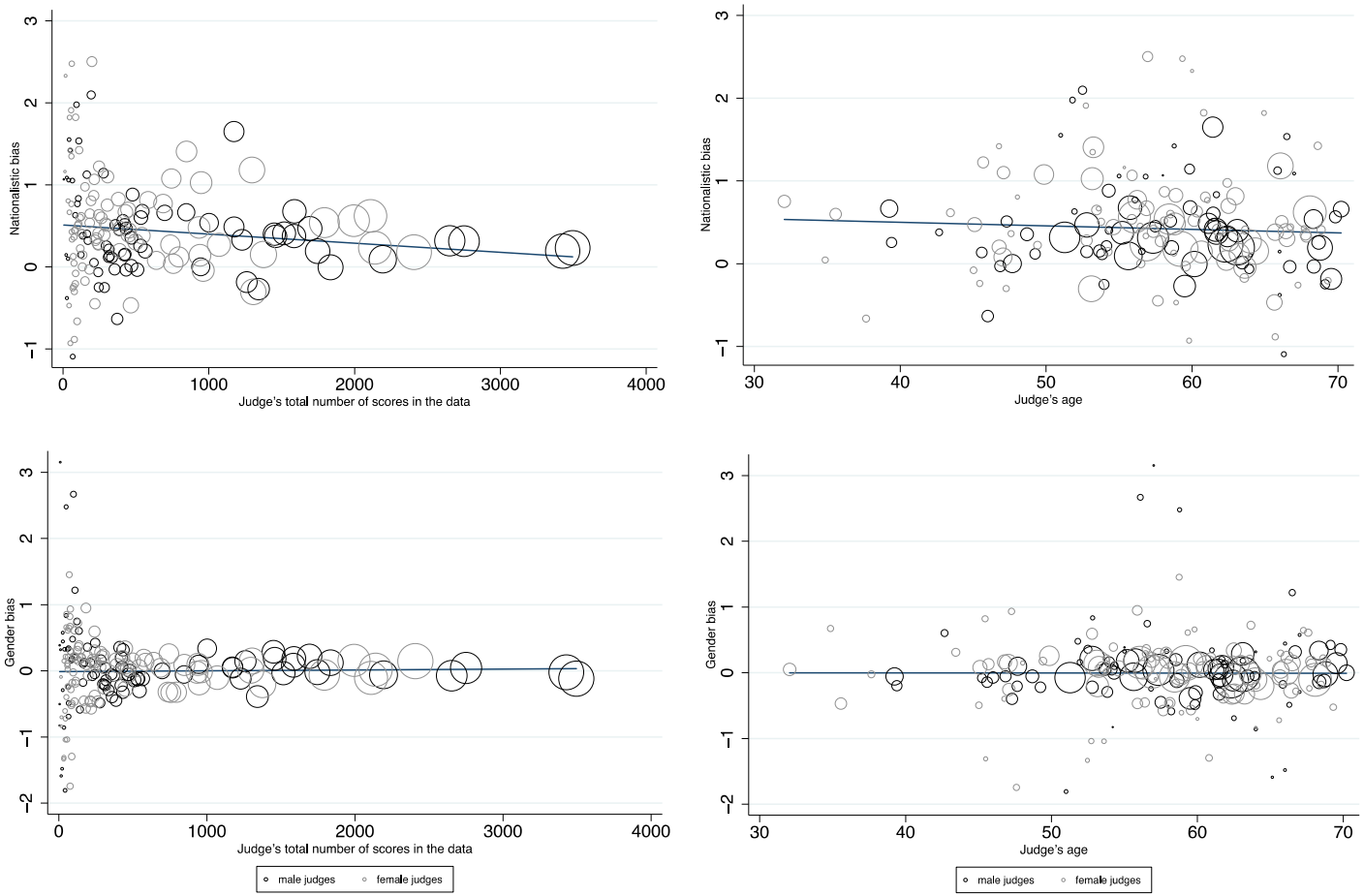


Fig. A5. *Correlation between Judge-Specific Bias and Judge Characteristics*

Notes. Each circle represents a judge-specific estimate of nationalistic bias (β_k from model (2)) or gender bias (α_k from model (3)). The size of the circles is proportional to the total number of observations in the data for each judge. The line shows the predicted value from a linear regression of the judge-specific bias on the variable indicated on the x-axis.

The Institute for International Economic Studies (IIES), Stockholm University

Additional Supporting Information may be found in the online version of this article:

Appendix B. Additional Tables and Figures.
Appendix C. Additional Analyses.
Appendix D. Additional Information about Dressage Events and Judges.

References

Abrevaya, J. and Hamermesh, D.S. (2012). 'Charity and favouritism in the field: Are female economists nicer (to each other)?', *Review of Economics and Statistics*, vol. 94(1), pp. 202-207.

Alesina, A., Devleeschauwer, A., Easterly,W., Kurlat, S. and Wacziarg, R. (2003). 'Fractionalization', *Journal of Economic Growth*, vol. 8(2), pp. 155-194.

Anwar, S., Bayer, P. and Hjalmarsson, R. (2012). 'The impact of jury race in criminal trials', *Quarterly Journal of Economics*, vol. 127(2), pp. 1017-1055.

Bagues, M.F. and Esteve-Volart, B. (2010). 'Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment', *Review of Economic Studies*, vol. 77(4), pp. 1301-1328.

Bagues, M.F., Sylos-Labini, M. and Zinovyeva, N. 'Does the gender composition of scientific committees matter?', *American Economic Review*, forthcoming.

Beck, T., Behr, P. and Madestam, A. (2014). 'Sex and credit: Is there a gender bias in lending?', European Banking Center Discussion Paper 2012-017.

Bernhard, H., Fehr, E. and Fischbacher, U. (2006). 'Group affiliation and altruistic norm enforcement', *American Economic Review*, vol. 96(2), pp. 217-221.

Blank, R.M. (1991). 'The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review', *American Economic Review*, vol. 81(5), pp. 1041-1067.

Booth, A. and Leigh, A. (2010). 'Do employers discriminate by gender? A field experiment in female-dominated occupations', *Economics Letters*, vol. 107(2), pp. 236-238.

Boring, A. (2017). 'Gender biases in student evaluations of teaching', *Journal of Public Economics*, vol. 145, pp. 27-41.

Bourhis, R.Y. and Gagnon, A. (2001). 'Social orientations in the minimal group paradigm', in (R. Brown and S. Gaertner, eds.), *Blackwell Handbook of Social Psychology: Intergroup Processes*, vol. 4, pp. 89-111, Oxford, UK: Blackwell.

Breda, T. and Hillion, M. (2016). 'Teaching accreditation exams reveal grading biases favour women in male-dominated disciplines in France', *Science*, vol. 353(6298), pp. 474-478.

Broder, I.E. (1993). 'Review of NSF economics proposals: Gender and institutional patterns', *American Economic Review*, vol. 83(4) pp. 964-970.

Charness, G., Rigotti, L. and Rustichini, A. (2007). 'Individual behaviour and group membership', *American Economic Review*, vol. 97(4), pp. 1340-1352.

Chen, R., and Chen, Y. (2011). 'The potential of social identity for equilibrium selection', *American Economic Review*, vol. 101(6), pp. 2562-2589.

Chen, Y., and Li, S.X. (2009). 'Group identity and social preferences', *American Economic Review*, vol. 99(1), pp. 431-457.

Dee, T.S. (2005). 'A teacher like me: Does race, ethnicity, or gender matter?', *American Economic Review*, vol. 95(2), pp. 158-165.

De Paola, M. and Scoppa, V. (2015). 'Gender discrimination and evaluators' gender: Evidence from Italian academia', *Economica*, vol. 82(325), pp. 162-188.

Duggan, M. and Levitt, S.D. (2002). 'Winning isn't everything: Corruption in sumo wrestling', *American Economic Review*, vol. 92(5), pp. 1594-1605.

Eckel, C. and Grossman, P. (2005). 'Managing diversity by creating team identity', *Journal of Economic Behaviour & Organization*, vol. 58(3), pp. 371-392.

Emerson, J.W., Seltzer, M. and Lin, D. (2009). 'Assessing judging bias: An example from the 2000 Olympic Games', *The American Statistician*, vol. 63(2), pp. 124-131.

Feld, J., Salamanca, N. and Hamermesh, D.S. (2016). 'Endophilia or exophobia: Beyond discrimination', *Economic Journal*, vol. 126(594), pp. 1503-1527.

Fisman, R., Paravisini, D. and Vig, V. 'Cultural proximity and loan outcomes', *American Economic Review*, forthcoming.

Goette, L., Huffman, D. and Meier, S. (2012). 'The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups', *American Economic Journal: Microeconomics*, vol. 4(1), pp. 101-115.

Goldin, C. and Rouse, C. (2000). 'Orchestrating impartiality: The impact of 'blind' auditions on female musicians', *American Economic Review*, vol. 90(4), pp. 715-741.

Guiso, L., Sapienza, P. and Zingales, L. (2009). 'Cultural biases in economic exchange', *Quarterly Journal of Economics*, vol. 124(3), pp. 1095-1131.

Guiso, L., Sapienza, P. and Zingales, L. (2004). 'Cultural biases in economic exchange', NBER Working Paper No. 1105.

Hargreaves Heap, S.P. and Zizzo, D.J. (2009). 'The value of groups', *American Economic Review*, vol. 99(1), pp. 295-323.

Leider, S., Möbius, M.M., Rosenblat, T. and Do, Q-A. (2009). 'Directed altruism and enforced reciprocity in social networks', *Quarterly Journal of Economics*, vol. 124(4), pp. 1815-1851.

Li, D. (2011). 'Gender bias in NIH peer review: Does it exist and does it matter?', unpublished manuscript, Kellogg School of Management, Northwestern University.

Maoz, Z. (2005). Dyadic MID Dataset (version 2.0): http://psfaculty.ucdavis.edu/zmaoz/dyadmid.html (last accessed 15 October 2016).

Mengel, F., Sauermann, J. and Zölitz, U. (2016). 'Gender bias in teaching evaluations', unpublished manuscript.

Mullen, B., Brown, R. and Smith, C. (1992). 'Ingroup bias as a function of salience, relevance, and status: An integration', *European Journal of Social Psychology*, vol. 22(2), pp. 103-122.

Parsons, C.A., Sulaeman, J. Yates, M.C. and Hamermesh, D.S. (2011). 'Strike three: Discrimination, incentives, and evaluation', *American Economic Review*, vol. 101(4), pp. 1410-1435.

Pettersson-Lidbom, P. and Priks, M. (2010). 'Behaviour under social pressure: Empty Italian stadiums and referee bias', *Economics Letters*, vol. 108(2), pp. 212-214.

Pope, B.R. and Pope, N.G. (2015). 'Own-Nationality Bias: Evidence from UEFA Champions League Football Referees', *Economic Inquiry*,vol. 53(2), pp. 1292-1304.

Price, J. and Wolfers. J. (2010). 'Racial discrimination among NBA referees', *Quarterly Journal of Economics*, vol. 125(4), pp. 1859-1887.

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*, vol. 106(15), pp. 6187-6191.

Shayo, M. and Zussman, A. (2011). 'Judicial ingroup bias in the shadow of terrorism', *Quarterly Journal of Economics*, vol. 126(3), pp. 1447-1484.

Sutter, M. (2009). 'Individual behaviour and group membership: Comment', *American Economic Review*, vol. 99(5), pp. 2247-2257.

Tajfel, H., Billig, M.G., Bundy, R.P. and Flament, C. (1971). 'Social categorization and intergroup behaviour', *European Journal of Social Psychology*, vol. 1(2), pp. 149-178.

Wennerås, C. and Wold, A. (1997). 'Nepotism and sexism in peer-review', *Nature*, vol. 387(6631), pp. 341-343.

Zitzewitz, E. (2006). 'Nationalism in winter sports judging and its lessons for organizational decision making', *Journal of Economics & Management Strategy*, vol. 15(1), pp. 67-99.