The effect of transparency on subjective evaluations: Evidence from competitive figure skating*

Ho, Chui Yee [†] Fang, Ximeng [‡]

September 1, 2023

Abstract

High-stakes decisions are often informed by aggregating views of multiple experts. The precision and accuracy of decisions then depends on the (strategic) incentives that these experts face. In this paper, we study the role of whether individual views are made transparent to the public or not, both theoretically and empirically. Specifically, we exploit a transparency reform in competitive figure skating and investigate its effect on performance evaluation by judge panels, using a difference-in-differences design. Prior to the reform, individual judges' scores in many events were kept anonymous, but scores were published openly in all events from the 2016-17 season onwards. We find that higher transparency leads to higher levels of implied consensus in subjective evaluations: the artistic scores awarded for a given performance (but not the more objective technical scores) become significantly less dispersed across judges. This consensus effect is stronger for high-profile competitions, and we find suggestive evidence that it is partly driven by higher precision of individual evaluations. However, we find no evidence that transparency reduces biases due to nationalistic favoritism. Our empirical results are consistent with a theoretical beauty-contest model in which transparency influences decision-making through increased conformity concerns.

JEL classification: D7, D82, L83

Keywords: transparency, subjective evaluations, committees, conformity, nationalistic favoritism, sports judges, difference-in-differences

[*] We would like to thank audiences at various conferences and workshops for their helpful comments and suggestions. Financial support by the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B07) is gratefully acknowledged.

[†] Bonn Graduate School of Economics, University of Bonn.

[‡] Saïd Business School, University of Oxford.

1. Introduction

High-stakes decisions and evaluations are often delegated to groups of experts, as opposed to a single individual. This includes, among many other examples, the recommendation and implementation of government policies through specialized committees, judicial rulings by panels of jurors or judges, hiring decisions in the labor market, and performance evaluation in professional sports. Drawing on the views of multiple evaluators can improve the accuracy and precision of the final decision or recommendation by collecting and aggregating information (à la Condorcet), while simultaneously mitigating the influence of idiosyncratic preferences and biases.

However, the effectiveness of aggregating multiple evaluations depends crucially on the institutional design and the (strategic) incentives generated by it. One important feature is whether the votes and opinions of each individual are made public or kept secret. On the one hand, higher transparency of the decision-making process allows the public to hold individual evaluators accountable, who may in turn try to stay more impartial and put in more effort in acquiring and communicating relevant information. On the other hand, transparency may expose evaluators to undesired influences (such as outside pressure), and it can also cause excessive conformity or conservatism, i.e., members becoming hesitant in expressing controversial opinions or deviating from a norm or consensus.[1] This may be particularly relevant in the absence of truly objective benchmarks for ex post validation. Thus, the effects of higher transparency on subjective decision-making can be theoretically ambiguous and nuanced (e.g., Levy, 2007; Gersbach and Hahn, 2012; Fehrler and Hughes, 2018; Mattozzi and Nakaguma, 2019; Fehrler and Janas, 2021). Yet, with a few notable exceptions (e.g., Meade and Stasavage, 2008; Benesch, Bütler and Hofer, 2018; Hansen, McMahon and Prat, 2018), causal evidence on the effects of transparency in real-world evaluation contexts remains scarce, mainly due to lack of suitable data and other empirical challenges.

In this paper, we study the effect of transparency on performance evaluation in the context of competitive figure skating. Figure skating is an inherently subjective sport, since the quality of an athlete's performance is partially derived from artistic aspects such as music interpretation and choreography. Hence, skaters' performances are independently evaluated by a panel of (typically nine) expert judges. Prior to the 2016-17 season, judges' scores in many competitions were published anonymously, meaning that only the distribution of scores and the identities of judges on the panel were known, but the two could not be linked to each other. In 2016, following allegations of biased evaluations due to nationalistic favoritism, a major transparency reform was implemented, so each judges' scores were

[1] The famous experiment by Asch (1951) is a classical example of how group conformity overrules reason. Similarly, it has been argued that the wisdom-of-crowds phenomenon may not hold when the aggregated judgements are not independent but exposed to social influence (Lorenz et al., 2011).

published openly from the 2016-17 season onwards. We examine the effects of this transparency reform on judges' performance evaluation behavior in a difference-in-differences design, using as control group a subset of events (Junior Grand Prix competitions) in which individual judges' scores were already published openly pre-reform.

This setting allows us to overcome several empirical challenges. First, we observe a large number of comparable decisions by professional evaluators in a high-stakes context, both under anonymous and transparent disclosure regimes. Second, the aggregation mechanism is common knowledge and we observe all inputs that contribute to the overall decision. Third, we can rule out joint deliberation and strategic agreements within the committee, as figure skating judges are not allowed to communicate with each other when awarding scores. Finally, the difference-in-differences setup allows us to control for general time trends unrelated to the reform, thus helping us to isolate the effect of higher transparency.

Individuals have generally been found to shift their behavior more towards the socially acceptable norm when (feeling) observed by others.[2] Accordingly, if judges want to appear competent and impartial in the public eye, then higher transparency could trigger judges' image and reputation concerns and thereby induce them to report more accurate evaluations (see, e.g., Suurmond, Swank and Visser, 2004; Bar-Isaac, 2012; Gersbach and Hahn, 2012; Hansen, McMahon and Prat, 2018; Mattozzi and Nakaguma, 2019; Swank and Visser, 2021). This may be of particular importance in the presence of significant subjective bias and favoritism in evaluation decisions, which has been well documented in figure skating and beyond.[3] However, there is no completely objective metric in figure skating against which judges' evaluation decisions can be validated against, i.e., the "accurate" score is never truly revealed — which is the very reason why performances are evaluated by a panel of expert judges in the first place. Thus, subjective performance evaluation includes elements of a credence good (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006). In such situations, a natural benchmark for evaluations of individual panel members is the comparison to evaluations by the other members.[4] This can create

[2] For example, students tend to reduce (visible) schooling investments when their rankings are revealed to their classmates (Bursztyn and Jensen, 2015), grocery store workers work harder when observed by more productive co-workers (Mas and Moretti, 2009), individuals are more likely to vote if they believe that their voting status would be revealed to their neighbors (Gerber, Green and Larimer, 2008).

[3] Systematic biases, especially in the form of nationalistic favoritism, has been documented in figure skating (Campbell and Galbraith, 1996; Zitzewitz, 2006; Lee, 2008; Litman and Stratmann, 2018) as well as in other professional sports where performance is evaluated by judge panels (see e.g. Sandberg, 2018). Relatedly, there is evidence for home team bias and racial bias in refereeing decisions (Garicano, Palacios-Huerta and Prendergast, 2005; Price and Wolfers, 2010; Parsons et al., 2011). Subjective biases are also prevalent in the evaluation of academic research (see, e.g., Li, 2017; Huber et al., 2022).

[4] Indeed, committee members are frequently evaluated by comparing them to their peers. This is based on the rationale that evaluations that are more accurate will generally be more strongly (positively) correlated with each other. In figure skating, large deviations from average scores can lead to disciplinary actions against judges.

strategic incentives for judges to become more “conformist”, i.e., to report scores that are closer to the scores that (they think) other judges will report. This can encourage higher individual effort to determine what would be objectively fair, but it could also lead to a loss of information value (Prendergast, 1993; Prat, 2005).

To explore the potential effects of transparency more formally, we present a theoretical model based on a beauty contest framework à la Morris and Shin (2002) with endogenous information acquisition. Judges are partially motivated by a truth-telling motive, but they also have a distortion motive due to subjective biases (such as favoritism toward compatriot athletes). Additionally, reputation-concerned judges have a conformity motive, i.e., they want to award scores that are similar to those of their fellow judges. We interpret higher transparency through the publishing of individual scores as an exogenous increase in this conformity motive. The model highlights three key mechanisms through which transparency can affect judge evaluation behavior. Firstly, judges exert higher effort to generate more precise signals, as a reduction in noise will generally lead to higher correlation of signals within the panel. Secondly, judges become more cautious and conservative in their scores, e.g. by anchoring towards a common prior, thus leading them to place lower weight on their private signal than they would under anonymous scoring. Lastly, transparency can induce judges to curb the expression of their idiosyncratic biases towards certain skaters; paradoxically, this may not lead to lower *aggregate* bias in the panel, as conformity concerns create the perverse incentive for judges to match the expected biases of other judges on the panel.

Several testable predictions arise. Above all, the model unambiguously predicts that the dispersion of scores across judges for a given performance will decrease after the transparency reform. This consensus effect is expected to be larger the more difficult it is to observe an objective score — implying in our context that conformity should be stronger for the artistic elements, rather than the technical elements of the performance —, the higher public attention on the performance is, and the stronger preconceived biases are (e.g., due to nationalistic favoritism). The model also predicts that, contrary to the aim of the reform, *aggregate* nationalistic bias will not necessarily decrease under greater transparency. To examine the effects of the transparency reform empirically, we analyze scores from almost 17,000 figure skating performances across 127 competitions organized by the International Skating Union (ISU) between 2013 and 2020. Our empirical identification strategy compares changes in the distribution of judge scores after the 2016 transparency reform between JGP (Junior Grand Prix) events, which were not affected by the reform, and Non-JGP events, which were.

Our empirical results are in line with the theoretical predictions. Importantly, we find that individual judges’ scores for a given performance become more similar to each other after the transparency reform takes effect. In particular, the dispersion of artistic scores within the judge panel drops sharply for Non-JGP events, relative to JGP events. The consensus effect in artistic scores is both statistically significant and quantitatively sizable — constituting approximately 9% of the pre-reform average and 29% of the pre-reform standard deviation of within-panel score dispersion — and it is mainly driven by the reduction of large outliers, so judges' scores become more tightly packed around the mean. It is also particularly pronounced for high-profile events, which arguably garner greater public attention, thus supporting the notion that the effects of transparency on judge evaluations are mediated by image and reputation concerns. However, we observe no consensus effect for the more objective technical score, which covers aspects like difficulty and execution of technical elements (jumps, spins, etc.). Moreover, there is no evidence that the reform led to a decrease in *aggregate* nationalistic bias, as measured by the average score advantage a skater receives when he or she has a compatriot judge on the panel. Although surprising given the reform's original intentions, this is consistent with our theoretical predictions.

Our theoretical framework highlights three mechanisms that can generate our empirical findings: higher effort, implicit coordination on common priors or signals, and conformity in biases. We find no evidence that judges give more similar scores the longer they have been evaluating together in the same panel, which speaks against implicit coordination through social learning. Furthermore, there is only weak evidence that the conformity effect is stronger for performances with a compatriot judge on the panel, and quantitatively it cannot fully explain the average decrease in score dispersion across judges. This suggests that a significant part of the consensus effect may be driven by more precise evaluations through higher effort or attention. To provide suggestive evidence for this, we analyze the sub-scores for different artistic components (e.g., choreography, music interpretation, transitions, ...) that sum up to the overall artistic score. We first document that within-judge consistency of sub-scores across artistic components could be interpreted as proxy for accuracy, as higher consistency is associated with other markers of evaluation quality at the individual judge level. Second, we document that consistency of artistic (but not technical) sub-scores increases significantly post-reform, which could thus be interpreted as marker for higher effort when awarding scores. As a robustness check, we verify that the transparency reform did not induce a different selection of judges into committees based on observable characteristics. Yet eventually, as we cannot determine an objective score for a performance without using the judge panel scores, we are not able to fully distinguish between these different mechanisms empirically.

Our paper contributes firstly to the literature on the consequences of transparency in committee decision-making. Theoretical models typically study how members' reputation concerns, i.e. their desire to appear competent, determine how they respond to transparency. Although transparency may under some circumstances induce anti-conformism to signal individual competence (Levy, 2007), committees may also have a preference for showing a united front in the public, in particular if true states cannot be observed ex post (Visser and Swank, 2007; Swank, Swank and Visser, 2008; Swank and Visser, 2021). Higher transparency can also lead to more pre-decision information acquisition (Gersbach and Hahn, 2012; Swank and Visser, 2021). One difference to our setting is that these theoretical papers typically study a binary decision, whereas scores in our setting are awarded on a scale and aggregated by averaging.[5] Empirical evidence on the effect of transparency on committee decision-making is relatively scarce. Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2019) provide laboratory evidence on the role of different transparency regimes on information aggregation in groups. With regard to real-world committees, several studies examine how monetary policy deliberations responded to a reform that resulted in transcripts of FOMC meetings being made public after Fall 1993. Meade and Stasavage (2008) find that members are less likely to voice disagreement with the Committee Chairman post-reform; using computational linguistics tools, Hansen, McMahon and Prat (2018) find that FOMC members tend to give more similar statements and engage less in back-and-forth dialogue post-reform, but also that especially rookie members seem to be better prepared with quantitative information on a diverse set of topics. Benesch, Bütler and Hofer (2018) study a transparency reform in the Upper House of the Swiss parliament and show that, post-reform, legislators exhibit greater party discipline. Though we also find a conformity effect, there are several noteworthy differences in our setting. Firstly, the report space in our setting is continuous, which allows for strategies that do not exist under a binary report space. Secondly, and more importantly, the lack of a deliberation or discussion stage in the current setup implies that the result we find is not due to (direct) coercion or coordination with other judges. Thus, this paper thus adds to this literature by demonstrating a conformity effect under greater transparency even in the absence of information exchange, thus providing stronger evidence for the way social image concerns can affect behavior of committee members.

A large number of previous studies have utilized large-scale publicly available data from professional sports contexts to investigate, among others, determinants of performance (e.g. Dohmen, 2008*a*; Lichter, Pestel and Sommer, 2017; Jiang, 2020), systematic decision errors (e.g. Pope and Schweitzer, 2011; Bruine de Bruin, 2006), gender differences (e.g. Böheim, Lackner and Wagner, 2020), as well as favoritism (e.g. Garicano, Palacios-Huerta and Prendergast, 2005; Zitzewitz, 2006; Sandberg, 2018; Fernando and George, 2021) and racial biases (e.g. Price and Wolfers, 2010; Parsons et al., 2011; Pope, Price and Wolfers, 2018). Two closely related papers to ours are by Zitzewitz (2014) and Lee (2008), who study a set of reforms in figure skating (following a vote trading scandal at the 2002 Winter Olympics) that in fact introduced the anonymous scoring regime that was eventually reversed in 2016. Zitzewitz (2014) finds a slight but statistically insignificant increase in the compatriot score advantage after the reform, and Lee (2008) finds an increase in the

[5] Rosar (2015) studies committee decision rules with continuous reporting and decision spaces and shows how this gives rise to incentives for strategic exaggeration.

standard deviation of judges' scores under anonymized publication. However, a number of other major reforms were implemented at that time, including an increase in the size of the judging panel and random dropping of judges' scores from the calculation of the final score, followed by another extensive series of reforms two years later. Our current setting using the 2016 reform allows for a cleaner attribution of changes in judge scoring behavior to increased transparency of judges' decisions, and our use of JGP events as control group in a difference-in-differences design further tightens the empirical identification by controlling for counterfactual time trends.

We also contribute to the literature studying whether changes in information structures could reduce discrimination. In recent years, a variety of reforms have been implemented at a large-scale (e.g. quotas, increased minority representation on selection committees, blind applications, pay transparency etc) to mixed results.[6] We provide a new empirical case study on the efficacy (or lack thereof) of a transparency-based method to counter favoritism/discrimination. Our results show that there is no evidence for any reduction in nationalistic favoritism following the publication of individual judge scores in figure skating. This could be due to several reasons. First, fairness norms might not be strong enough or offset by opposing loyalty norms induced by judges' home audience. Second, the group structure of committees could interact with conformity concerns, so that judges aim to give more similar scores to their peers by matching their biases, or alternatively, that the non-compatriot judges might skew their scores slightly upwards when one of their peers has the same nationality as the skater.[7] Third, the bias-correcting properties of aggregating multiple votes reduces the scope for reducing the aggregate bias.

The remainder of the paper is organized as follows. Section 2 gives a brief overview of our empirical context. In Section 3, we discuss how transparency can lead to changes in behavior through the lens of a theoretical model. We describe our data and provide summary statistics in Section 4. The empirical strategy is outlined in Section 5. In Section 6, we present our main empirical results, and Section 7 shows additional results to explore the underlying mechanisms. Section 8 concludes.

[6] See, e.g., Bertrand et al. (2018); Maida and Weber (2019) for evidence on quotas, Bagues and Esteve-Volart (2010); Bagues, Sylos-Labini and Zinovyeva (2017) for evidence on the effectiveness of gender representation on selection committees, Behaghel, Crépon and Le Barbanchon (2015); Krause, Rinne and Zimmermann (2012) on blind applications, Baker et al. (2019); Mas (2017) on pay transparency.

[7] Bagues and Esteve-Volart (2010) also hint at strategic dependencies between committee members leading to worse outcomes for female candidates paired with academic committees with greater female representation, as male committee members became less favorable when there were more female members on the committee.

2. Context

Figure skating is a sport in which athletes (individuals or pairs) skate on ice and perform a choreographed sequence of jumps, spins, and dance moves to a musical track. There are four main disciplines in figure skating: Men's Singles, Women's Singles, Pairs Skating, and Ice Dance. In this paper, we focus on official international events recognized by the International Skating Union (ISU). Some of the most prestigious ISU events include the World Championships, the Grand Prix Series and Finals, and the quadrennial Olympics Winter Games. Each event typically consists of four competitions, one for each discipline. Within each competition, skaters skate twice, once in the Short Program and once in the Long Program. The skater's final placement in the competition is determined by the sum of total scores in each program.

2.1. Scoring of figure skating performances

Within the ISU Judging System, skaters are evaluated by a panel of (typically) 9 judges, who watch the performance and award scores to indicate its technical and artistic quality. Judges are not allowed to confer with each other while grading the performance.Scores consist of two main parts: the Technical Elements Score (TES), which evaluates the difficulty and execution of technical elements, and the Program Component Score (PCS), which evaluates the artistic value of the performance. The Total Score (TS) for a skating performance is given by the sum of the TES and the PCS, minus any potential deductions (e.g., due to rule violations). Throughout the paper, we will often refer to the TES as the "technical score" and to the PCS as the "artistic score".

The TES is determined as follows. Skaters perform a number of technical elements (jumps, spins, etc.) in their performance, and each element receives a score from the judge panel. This score is computed based on the Base Value, which increase in the difficulty level of the element, and the Grade of Execution (GoE), which is assigned by each member of the judge panel and indicates how cleanly the element was executed.[8] This GoE is then scaled according to the difficulty of the element and added to its Base Value, with more difficult technical elements receiving higher GoE scaling factor. To hinder manipulation and reduce the impact of outliers, the highest and lowest GoEs for each technical element in the judge panel are dropped. The overall TES of a performance is obtained by calculating the (trimmed) average scores for all technical element across judges and summing them up.

In contrast to the TES, the artistic scores that determine the PCS are awarded after the end of the performance. Each judge assigns a score to the artistic components of performance, which include the interpretation of music, skating skills, transitions between

[8] The GoE ranges between -3 and +3, with increments of 1. From the 2018-19 season onwards, the range of the GoE was increased, to span from -5 to +5.

technical elements, composition, and performance. Each component can be marked on a range from 0.25 to 10 in quarter-point increments. Again, the highest and lowest scores in the judge panel for each component are dropped. The PCS is obtained by calculating the (trimmed) average scores for all components across judges and summing them up.

2.2. Transparency reform in 2016

Each season, there are around 20 ISU events, including the European Championships, Four Continents Championships, World Championships, Olympics Winter Games, the Grand Prix Series and Final, the Junior World Championships, and Junior Grand Prix (JGP) Series and Final. After each event, the ISU publishes detailed scoring information for all performances, including the individual judge scores that make up the final score, its official website. Prior to the 2016-17 season, with the exception of Junior Grand Prix (JGP) Series events, these individual scores were published anonymously. That is, while the identities of the judges on the panel were known, the individual scores are published in random order, so that they cannot be linked to an individual judge.[9]

This lack of transparency meant that judges could not be held accountable for their decisions, which led to accusations of biased judging by the public. Such allegations came to a head with the scoring of the 2014 Olympics Ladies competition, where Russian competitor Alina Zagitova was awarded gold ahead of the South Korean competitor Kim Yu-Na. Indeed, public outrage over the scoring reached such a point that the International Skating Union (ISU) considered abolishing judge anonymity in their General Meeting in 2014. While the proposal failed narrowly, it was brought up once again two years later (in 2016) and passed, so that from 2016-17 onwards, judges' scores from all competitions were published openly. Though other reforms were implemented at the 2016-17 meeting, these reforms were not explicitly aimed at reducing nationalistic judging, and mostly affect both JGP (Control) and Non-JGP (Treatment) events.[10]

Because JGP events already published scores openly prior to the transparency reform, they were unaffected by the reform and thus serve as a control group. JGP events follow the same scoring format and criteria as Non-JGP events and, to a certain extent, share the same pool of judges as Non-JGP events— over the study period of 2013-2020, half of the judges have judged in at least one JGP event and Non-JGP event. The core difference between these two groups of events lies in the level of prestige and exclusivity. JGP events are typically less prestigious and exclusive than Non-JGP events, so that scores from JGP events tend to be lower.

[9] See Figure A.1 for an example of a published score sheet.

[10] Other reforms are mostly concerned with changes in required technical elements and updated scoring guidelines, which are typically implemented every two years (when a General Meeting is held). A few rule changes are specific to Senior events; however, these are mostly specific to the technical elements.

3. Theoretical Framework

The main consequence of the transparency reform is that individual judges' evaluations become perfectly observable, with the aim of encouraging more accurate and less biased judge evaluations through reputational incentives. Thus, the idea is that career-concerned judges will want to appear competent and impartial in the face of public scrutiny. However, there is often no truly objective yardstick against which an individual judge's evaluation accuracy can be compared against. This is clearly the case in the context of competitive figure skating, as the subjective nature of the sport is the very reason why athletes' performances are evaluated by aggregating multiple individual scores from panel of expert judges.

A natural and intuitive approach to evaluate the marking accuracy and impartiality of individual judges is to compare their scores against the scores awarded by the other expert judges on the panel (Heiniger and Mercier, 2021). Outlier judges who express very different opinions from those of their peers may be perceived as being incompetent, inattentive, or biased, whereas judges who are close to the median might be perceived as competent and impartial.Therefore, the transparency reform plausibly generates stronger incentives for judges to report scores that are more similar to those of others. Note that it is not possible (and not allowed) for judges to deliberate together or coordinate their scores, but judges could potentially react to transparency by exerting more effort into marking accurately, by curbing their biases toward certain skaters (e.g., of the same nationality), or by anchoring conservatively towards a common prior.

To formalize these intuitions and to derive predictions for how transparency could affect the distribution of scores within the judge panel, we present a theoretical model of judges' performance evaluation behavior that is based on the well-studied beauty contest framework introduced by Morris and Shin (2002), and extended by Colombo and Femminis (2008) to incorporate costly information acquisition.

3.1. Basic setup

Skater i performs in a competition. Judges $j = 1, ..., N$ sit on the panel and evaluate the quality of the performance by each reporting a score π_{ji} without joint deliberation. These individual scores $\pi_{1i}, ...\pi_{Ni}$ are then aggregated to an overall average score $\pi_i = \frac{1}{N}\sum_j \pi_{ji}$. For simplicity, we abstract from the trimming of the highest and lowest scores.

The common prior of performance quality θ_i for skater i follows a normal distribution with mean μ_i and (non-zero) variance σ_i^2. Judges may reasonably have different priors about, e.g., a consistently world-class skater compared to a capricious rookie, so both μ_i and σ_i can differ across skaters. As there is a strong artistic aspect to figure skating and thus no simple objective criterion for evaluating a performance, the "true" realized quality θ_i is imperfectly observable ex post. However, by watching the performance, each judge receives a private signal of the performance quality:

$$x_{ji} = \theta_i + \varepsilon_{ji}\,, \tag{1}$$

which can be thought of as reflecting the judge's own personal assessment.[11] The signal is unbiased but contains an idiosyncratic noise term ε_{ji} that is independent of θ_i and that follows a normal distribution with mean 0 and variance σ_i^2/τ_{ji}, where $\tau_{ji} \in (1, \infty)$ denotes the precision of judge j's signal for skater i. We assume that the private signal after oberving the performance is always more informative than the prior ($\tau_{ji} > 1$), but never so informative that θ_i is perfectly observed ($\tau_{ji} < \infty$). This offers a rationale for assigning final scores by aggregating the (independent) opinions of multiple judges in order to reduce the influence of idiosyncratic tendencies and judgement errors. However, ε_{ji} can be heteroscedastic. For example, an experienced and attentive judge may be able to evaluate the quality of a performance more reliably than a judge who is inexperienced or inattentive. Similarly, a performance that is excellent all around is arguably easier to evaluate than a mediocre performance with highs and lows.

3.2. Simplified model

To build intuition, we will first present a stripped-down version of our model in which judges behave non-strategically and in which signal precision τ_{ji} is given exogenously. We assume that judges are partially motivated to give a genuinely accurate assessment of the performance quality when reporting their scores, but that they can additionally be biased towards rewarding systematically higher or lower scores to skater i. This bias may reflect favoritism, e.g. due to same nationality or a preferred skating style (Zitzewitz, 2006; Litman and Stratmann, 2018), but it could in principle also reflect stable differences in judges' general strictness or leniency, if the bias is invariant to the skater's identity. We model these two elements through the following payoff function:

$$u_j(\pi_{ji}, b_{ji}, \theta_i) \;=\; -\,(\pi_{ji} - \theta_i - b_{ji})^2\,. \tag{2}$$

b_{ji} is the (fixed) bias of judge j towards skater i. Judges choose π_{ji} to maximize their expected utility. The quadratic loss formulation leads to a classical signal extraction problem, and the optimal non-strategic report $\tilde{\pi}_{ji}$ can be obtained using Bayes' rule:

$$\tilde{\pi}_{ji} \;=\; E[\theta_i | x_{ji}, y_i] \,+\, b_{ji} \;=\; \frac{1}{1+\tau_{ji}}\,\mu_i \,+\, \frac{\tau_{ji}}{1+\tau_{ji}}\,x_{ji} \,+\, b_{ji}\,. \tag{3}$$

[11] We simplify the Morris and Shin (2002) framework by not including a public signal y_i that is the main focus of their paper and of much of the literature it spurred. However, the skater-specific prior with mean μ_i and variance σ_i^2 could be interpreted implicitly as the interim posterior distribution conditional on public information about ex ante obervable characteristics of skater i, such as their previous performance scores.

The first component $E[\theta_i|x_{ji}, y_i]$ is a linear combination of the private signal x_{ji} and the common posterior μ_i and represents the actual posterior belief about performance quality θ_i that the judge forms. The more accurately a judge is able evaluate the performance, i.e. the higher τ_{ji}, the more weight will be put on his or her actual signal. The second component b_{ji} creates a distortion in the reported score due to the judge's bias towards skater i. Depending on how the biases are distributed across judges in the panel, they may not completely average out when scores are aggregated, so some skaters may have an unfair advantage compared to others, if it so happens that the panel is tilted in favor of them, e.g., if a compatriot judge sits on the panel.

Assuming homogenous precision $\tau_{ji} = \tau_i$ for all judges, the expectation and variance of scores across judges in the panel conditional on the performance θ_i are

$$E[\tilde{\pi}_{ji}|\theta_i] = \theta_i + \frac{1}{1+\tau_i}(\mu_i - \theta_i) + E[b_{ji}], \tag{4}$$

$$Var[\tilde{\pi}_{ji}|\theta_i] = \frac{\tau_i}{(1+\tau_i)^2}\sigma_i^2 + Var[b_{ji}]. \tag{5}$$

The overall score can be ex post biased from two sources. First, the reported scores are conservative, i.e., slanted towards the common prior expectation μ_i, because judges can only observe θ_i with noise. Hence, hypothetically, the identical performance delivered by a famous world-class skater may be awarded a higher score than if delivered by an unknown rookie skater — this is sometimes referred to as the Matthew effect (Merton, 1968; Kim and King, 2014; Huber et al., 2022). Second, a skater will receive systematically higher or lower scores if there is asymmetry in judges' biases, for example if one judge exhibits strong nationalistic favoritism and the other judges in the panel are unbiased. While public focus often lies on bias and favoritism, a reduction in noise can be equally important in ensuring the validity of a decision making process (Kahneman, Sobony and Sunstein, 2021). The expected variance of scores decreases with higher signal precision τ_i and with lower bias heterogeneity $Var[b_{ji}]$ across judges.

3.3. Full model

Our full model extends the non-strategic setup from above with two elements. First, judges are reputation-concerned, meaning that they want to appear competent in the way they award scores to a skating performance. As performance quality is not perfectly observable even ex post, especially with regard to the more artistic aspects, one straightforward way to evaluate a judges' score is to compare it to the score of other judges. Therefore, we model image concerns in a way that they lead to a motive for conforming with other judges, i.e. by not deviating too far from their scores. Second, we allow judges to endogenously adjust their signal precision τ_{ji} through costly information acquisition, which could be interpreted as level of effort or attentiveness when observing the performance.The judge's payoff function is

$$u_j(\pi_i, \tau_{ji}, \theta_i) = -\left(\pi_{ji} - \theta_i - b_{ji}\right)^2 \; - \; \eta \left(\pi_{ji} - \tfrac{1}{N-1} \sum_{l \neq j} \pi_{li}\right)^2 \; - C(\tau_{ji}) \, , \qquad (6)$$

where $\eta \in (0, 1)$ captures the strength of the conformity motive relative to the truthfulness motive, and $C(\tau_{ji})$ is the effort cost necessary to achieve precision level τ_{ji}. Following Colombo and Femminis (2008), we assume a linear cost function $C(\tau_{ji}) = c\tau_{ji}$. The unit "price" of precision is $c \in (0, \bar{c})$, with upper limit $\bar{c} = \frac{\sigma_i^2}{4\,(1+\eta)}$ to ensure that agents choose signal precisions τ_{ji} that are not implausibly low.[12] Note that there is now a strategic aspect to reporting behavior, since judge j's expected utility depends on the scores of the other judges, and vice versa. As a solution concept, we compute the symmetric Bayesian Nash equilibrium, in which each judge makes inferences about the distribution of other judges' signals based on her own signal and then awards her optimal scores in response to other judges' reporting strategy. The individual rationality condition requires that for all $j = 1, ..., N$ and $l \neq j$,

$$\begin{aligned} \pi_{ji} &= \frac{1}{1+\eta} \left(E[\theta_i | x_{ji}, y_i] \; + \; b_{ji}\right) + \; \frac{\eta}{1+\eta} \, E[\pi_{li} | x_{ji}, y_i] \\ &= \frac{1}{1+\eta} \, \tilde{\pi}_{ji} \; + \; \frac{\eta}{1+\eta} \, E[\pi_{li} | x_{ji}, y_i] \, . \end{aligned} \qquad (7)$$

As already observed by Morris and Shin (2002), a symmetric equlibrium implies that we can plug in the best response π_{li} from equation (7) for all $l \neq j$, leading to a feedback loop of higher-order beliefs that converges to a unique social equilibrium in which every judge j reports

$$\pi_{ji} = \frac{1+\eta}{1+\eta+\tau_{ji}} \, \mu_i \; + \; \frac{\tau_{ji}}{1+\eta+\tau_{ji}} \, x_{ji} \; + \; \frac{1}{1+\eta} \, b_{ji} \; + \; \frac{\eta}{1+\eta} \, E[b_i] \, . \qquad (8)$$

This equilibrium condition has to be true regardless of the level of precision τ_{ji} that judges choose. Holding constant τ_{ji}, the optimal strategic report π_{ji} is more conservative than the non-strategic report $\tilde{\pi}_{ji}$, i.e., it is attenuated more strongly towards the common prior expectation μ_i. Hence, it resembles a tacit coordination of judges to deviate from their true posterior beliefs of performance quality and move their scores closer towards an uncontroversial benchmark. Interestingly, the desire to appear more in line with other judges also leads to conformity in biases, as judges now realign their bias partially towards the expected bias $E[b_i]$.

[12] As we will later see, this condition on c implies that $\tau_{ji} > 1 + \eta$ and ensures that judges will always place more weight on their private signal than on the common posterior when reporting their score, which is arguably a reasonable assumption. This also ensures that the variance of scores always decreases in signal precision, because when judges placed a higher weight on the common posterior than the private signal, scores would become very uniform.

Next, we need to find the equilibrium level of effort τ_{ji}. Let all judges $l \neq j$ follow the same strategy, with report π_{ji} from equation (8) and homogeneous effort level $\tau_{li} = \tau_i$. Judge j takes this as given and and seeks to determine his or her individual effort level τ_{ji}. Adapting the results from Colombo and Femminis (2008), the optimal signal precision for all judges j in a symmetric equilibrium can be shown to be

$$\tau_{ji} = \tau_i = \sqrt{1+\eta} \cdot \frac{\sigma_i}{\sqrt{c}} - (1+\eta)\,. \tag{9}$$

Notice that this term is increasing in the conformity concern η for all $c \in (0, \bar{c}]$. Hence, transparency can be used as reputational incentive mechanism for inducing higher judge effort when evaluating skater performances.

Conditional on θ_i, the expectation and variance of performance scores across judges look as follows when taking into account conformity concerns and endogenous signal precision:

$$\begin{aligned} E[\pi_{ji}|\theta_i] &= \theta_i + \frac{1+\eta}{1+\eta+\tau_i}\,(\mu_i - \theta_i) + E[b_{ji}] \\ &= \theta_i + \frac{\sqrt{(1+\eta)\,c}}{\sigma_i}\,(\mu_i - \theta_i) + E[b_{ji}]\,, \end{aligned} \tag{10}$$

$$\begin{aligned} Var[\pi_{ji}|\theta_i] &= \frac{\tau_i}{(1+\eta+\tau_i)^2}\,\sigma_i^2 + \frac{1}{(1+\eta)^2}\,Var[b_{ji}] \\ &= \frac{\sqrt{c}\,\sigma_i}{2\,(1+\eta)^{\frac{3}{2}}} - c + \frac{1}{(1+\eta)^2}\,Var[b_{ji}]\,. \end{aligned} \tag{11}$$

The distribution of judge scores still follows similar properties as in the simple model. Judges' scores exhibit conservatism towards the prior expectation and scores are further distorted through the average bias toward skater i in the judge panel. The more precisely judges can observe the performance quality, the less conservative and the less noisy the scores become. On top of that, the full model also allows us to study how the score distribution is affected by the conformity motive μ, which is arguably affected by whether judging is transparent or anonymous. In the following, we will use the results in equations (10) and (11) to derive testable predictions for the effects of the transparency reform.

3.4. Predicted effects of the transparency reform

Under anonymous scoring, the public cannot observe which judge gave which score. Hence, judges do not have to worry much about appearing incompetent or biased when the score they award is discrepant from the other judges' scores. In contrast, when scoring becomes transparent, judges may start worrying more about their social image and their desire to appear competent. In our model, we therefore interpret scoring under transparency as an increase in η compared to scoring under anonymity. Conducting comparative statics with regard to η then allows us to derive a number of testable predictions for how the transparency reform affects judges' scores, which we list below.

(1) Lower score dispersion for a given performance. — If transparency leads to stronger conformity concerns, the variance of scores across judges in the panel for a given performance decreases:

$$\frac{\partial}{\partial \eta} Var[\pi_{ji}|\theta_i] \, < \, 0 \, . \tag{12}$$

There are three reasons for this lower score dispersion. First, stronger conformity concerns result in scores that are more conservative in the sense that they are attenuated towards the common posterior z_i, which means that judges place less weight on their idiosyncratic information. Second, increasing effort in η leads to less noise in judges' private signals. Third, dispersion can further decrease due to judges adjusting their individual biases more towards the average bias in the panel, which implies that the impact of transparency would be stronger if $Var[b_{ji}]$ is high, meaning that judges are very polarized in their biases towards a skater.

(2) Effect on score dispersion increases in subjectivity. — Skaters are evaluated both on the technical aspects and the artistic aspects of their performance. The latter is arguably much more subjective than the former, which implies that judges may have a harder time trying to award the artistic score as accurately as possible. We therefore look at another comparative static, which is how the effect of transparency on dispersion of scores is affected by an increase in the level of subjectivity/noise σ_i when judging performance quality. It is straightforward to show that

$$\frac{\partial^2}{\partial \eta \, \partial \sigma_i} Var[\pi_{ji}|\theta_i] < \, 0 \, . \tag{13}$$

This implies that the reduction in score dispersion in prediction (1) is more pronounced if objective performance evaluation is more difficult. In particular, we would expect to see a larger reduction in dispersion for the artistic score than for the technical score.

Note that the same would hold if we replaced σ_i with the cost of information acquisition c_i. Further rationales for expecting smaller effects for the technical score is that conformity to other judges may play less of a role (i.e. η is lower), because its relative objectivity makes it more important for reputation-concerned judges to give their most accurate assessment, or because technical scores are awarded almost instantaneously and judges may not have time to consider other judges' behavior.

(3) No decrease in aggregate bias. — Perhaps surprisingly, our model suggests that, on average, higher transparency may leave the *aggregate* bias $B_i = \sum_j b_{ji}$ of the panel towards skater i unchanged, as the bias component in equation (10) is invariant to

η:

$$\frac{\partial^2}{\partial\eta\,\partial E[b_{ji}]} E[\pi_{ji}|\theta_i] = \,0\,. \tag{14}$$

The reason is that with conformity concerns, judges also incorporate beliefs about other judges' biases $E[b_i]$ in order to match their scores more closely. This prediction is consistent with the results in Sandberg (2018), who finds that judges in dressage competitions favor athletes of the same nationality as other judges on the panel. In our context, one may therefore also expect conformity effects to be particularly strong when judge biases can be easily inferred, such as when there are matching nationalities.

3.5. Further potential channels of transparency

Transparency may also affect judge behavior through other mechanisms that are not explicitly included in our model. In the following, we will briefly discuss some of these mechanisms and how they may affect our theoretical predictions.

Appealing to the home constituency. Public monitoring generally induces individuals to behave more in accordance to prevailing norms and expectations, but these might not necessarily encourage impartiality. For example, audiences in the judge's home countries and the national federation that appointed the judge may in fact expect him or her to favor compatriot skaters and discriminate against rival skaters (Zitzewitz, 2006).[13] If this was the case, we would expect transparency to lead to an increase in nationalistic judging and an increase in score dispersion for performances with a compatriot judge on the panel, contrary to the predictions of our model.

Exaggeration and counterexaggeration. When there is a potentially biased judge on the panel, other judges can in fact react to this strategically by biasing their scores in the opposite direction if they have fairness concerns for the aggregate score awarded to skaters (Li, Rosen and Suen, 2001; Rausser, Simon and Zhao, 2015). Transparency could potentially break such feedback loops of bias and counterbias, which would also predict a decrease in score dipersion for a given performance, though mostly concentrated on performances where the presumed biases are particularly strong, e.g. when there is a compatriot judge on the panel. Note, however, that some previous studies on the behavior of sports judge panels find that non-compatriot judges may in fact move their scores closer

[13] Dohmen (2008*b*), for instance, finds that football referees exhibit home team favoritism, in particular when the physical distance of the public crowd to the field is smaller, and when the crowd consists of supporters of the home team. Benesch, Bütler and Hofer (2018) find greater party discipline after the transparency reform in the Swiss Upper House, even though this is not necessarily in line with the preferences of the median cantonal voter. Stasavage (2007) finds that in a model with biased and unbiased experts, unbiased experts only vote truthfully under public voting if reputational concerns are sufficiently weak.

towards those of the compatriot judge instead of the opposite (Zitzewitz, 2006; Sandberg, 2018).

Vote trading and rigging. Transparency can also facilitate corruption, e.g. by rigging or vote trading, because potential bribers can now verify whether the bribed judge actually followed through, and colluding judges can better monitor each others' behavior and implement repeated game strategies.[14] However, assuming that vote trading strategies need to be sophisticated enough that they are not easily detectable, it is difficult to predict how observed scoring patterns would be affected. Since collusion and cheating are risky endeavors with uncertain success chances, given the limited impact of individual judges, it seems unlikely that this would cause strong universal changes in observed judging behavior.

4. Data and Descriptive Statistics

To study how the 2016 transparency reform affected performance scoring by judges, we obtain from the ISU website information on skaters' performances at all official ISU competitions from the 2013-14 season to the 2019-20 season. Thus, our sample includes three pre-reform seasons under the anonymous scoring regime and four post-reform seasons under the transparency regime.[15] This information includes all scores awarded by judges on the panel towards each technical element and artistic program component of the performance, as well as the identities and nationalities of the skater and of the judges.

In total, our sample comprises 16,821 skating performances by 1,905 different skaters across 127 events. A figure skating event (e.g., 2018 Winter Olympics) can typically be further broken down into four competitions, one in each of the four disciplines (Men's Singles, Women's Singles, Pairs Skating, Ice Dance), and two rounds per competition (Short Program and Free Skating).[16] Within each round, the judge panel stays constant,

[14] In fact, anonymous voting was first introduced by the ISU in 2002 precisely in response to a vote trading scandal at the Salt Lake City Olympics, where a French judge admitted (though later recanted) to having been pressured by her national federation to rank the Russian pair first in the pairs' competition, in exchange for higher votes to a French couple that would perform in the ice dance competition a few days later.

[15] Though data is available until the 2005-06 season, the main presented results are restricted to observations from the 2013-14 season onwards. This is firstly due to a number of changes in event formats in the 2010-11 and 2011-12 seasons (e.g. the Compulsory Dance and Original Dance segments were replaced with the Short Dance segment; instead of holding a Preliminary Qualification Round in Senior events, qualifications were done based on scores from the Short Program after the 2011-12 season.), so that it is not possible to control for discipline × segment. Secondly, JGP (Control) skaters typically do not have long careers, so these skaters are no longer in the dataset after a few years; results with skater FEs are mainly identified from performances close to the reform period. Results using the full dataset (without skater FEs or discipline × segment controls) are presented in the Appendix.

[16] Note that the number of rounds is not 8 times the number of events in our sample, because some events hold more than one competitions per discipline, whereas some (JGP) events do not hold a competition for each discipline.

Table 1: Number of Observations

		JGP (control)		Non-JGP (treated)	
	full sample	pre-reform	post-reform	pre-reform	post-reform
# Performances	16821	3103	4340	3994	5384
# Events	127	21	28	34	44
# Rounds	1028	152	200	292	384
# Skaters	1905	711	954	617	730
# Judges	563	333	379	323	338

Notes. This table shows the number of observations in our sample, split by JGP events and Non-JGP events before and after the 2016 reform, respectively. An event typically consists of 4 competitions, one for each discipline (Men's Singles, Women's Singles, Pairs Skating, Ice Dance), and each competition consists of 2 rounds (Short Program and Free Skating). However, some JGP events do not include a Pairs Skating competition, and some other events hold more than one competition per discipline. We exclude 520 performances for which the panel included fewer than 9 judges.

so all skaters performing in the same round are evaluated by the same judges. Table 1 further breaks our sample down into observation categories according to our difference-in-differences identification strategy. We observe a comparable sample of performances in both treated Non-JGP events and untreated JGP events, although the number of observations is slightly lower for JGP events. Furthermore, as we include four post-reform and three pre-reform seasons, we have slightly more observations under transparency than under anonymity. We restrict the dataset to performances from competitions where there was a full panel of 9 judges.[17]

Table 2 presents descriptive statistics for the performance scores in our sample. The average Program Component Score (PCS), i.e., the artistic score, is about 38.08 over all performances, and the average Technical Elements Score (TES) is about 39.16. The average Total Score is somewhat lower than the sum of both, as skaters are sometimes punished with score deductions for rule violation. In general, scores in JGP events tend to be somewhat lower compared to Non-JGP events, reflecting the lower level of prestige and hence lower average quality of performances. Furthermore, there seems to be an upward time trend for all event types, so average post-reform scores tend to be higher the average pre-reform scores.

Judges are not unanimous in their evaluation decisions. As measure of disagreement about a performance in the panel we calculate the within-panel standard deviation (Panel SD), i.e., the score dispersion across judges for any given performance: $\sigma_p = \sqrt{\frac{1}{9}\sum_{j=1}^{9}(\pi_{pj} - \bar{\pi}_p)^2}$, where π_{pj} is the score awarded by judge j towards performance p. From Table 2, we can

[17] Due to budget constraints, some competitions (typically JGP) have panels with fewer than 9 judges. Nonetheless, such panels are uncommon, consisting only of 520 performances. Including these observations does not lead to in any significant changes in results.

Table 2: Descriptive Statistics

		JGP (control)		Non-JGP (treated)	
	full sample	pre-reform	post-reform	pre-reform	post-reform
Program Component Score (PCS)					
Average score	38.08	30.95	33.09	41.06	44.00
Mean Panel SD	1.75	1.83	1.84	1.78	1.62
Compatriot mean	40.46	31.74	34.50	43.06	46.24
Technical Elements Score (TES)					
Average score	39.16	31.09	33.72	42.08	46.04
Mean Panel SD	1.33	1.03	1.18	1.40	1.56
Compatriot mean	41.61	32.02	35.56	43.78	48.16
Total Score					
Average score	76.75	61.42	66.19	82.73	89.67
Mean Panel SD	3.13	2.98	3.13	3.20	3.17
Compatriot mean	81.62	63.18	69.52	86.42	94.04
% Compatriot	61	54	52	66	68

Notes. This table shows the number of observations in our sample, split by JGP events and Non-JGP events before and after the 2016 reform, respectively.

see that the mean Panel SD is about 1.75 for the PCS and 1.33 for the TES over all performances, reflecting the subjective nature of the sport. Another way to illustrate the magnitude of dispersion is by the calculating the gap between the highest and the lowest score in the judge panel for the same performance: this gap is 5.73 points for the PCS and 4.31 points for the TES. Notice that there is generally less disagreement on the more objective technical score compared to the artistic score. Notice also that the mean Panel SD of artistic scores drops from 1.78 to 1.62 in Non-JGP events after the transparency reform was introduced, whereas it stayed nearly unchanged in JGP events that were not affected by the reform.

Finally, Table 2 also shows the mean scores for compatriot performances, defined as performances for which there is at least one judge on the panel who has the same nationality as the skater. This is true for about 61% of performances in our full sample. In general, we observe that compatriot performances tend to be receive higher score relative to non-compatriot performances. Naturally, this compatriot score gap alone is no evidence for nationalistic favoritism. Countries that are traditionally strong in figure skating (such as China, Russia, USA, and Japan) are also overrepresented on judge panels, since judges are often former competitive figure skaters themselves, so a positive correlation between compatriot performances and scores is to be expected.

5. Empirical Strategy

5.1. Identification

We use a difference-in-differences approach to empirically identify the effects of the transparency reform on judges' performance evaluation behavior, using perfomances in JGP events as control group, since deanonymized scores were already published before the 2016 reform for these events. The main identification assumption is that performance scores in treated Non-JGP events and in untreated JGP events would have followed the same counterfactual time trend in absence of the transparency reform. While JGP events are notably less prestigious than Non-JGP events, any level differences in performance score statistics between these events are not problematic as long as the common trends assumption holds. Moreover, we need to assume that the reform does not affect skaters' performance per se (in an unobservable way), but only the way judges award scores for these performances. This seems plausible given that for skaters, nothing changes about how and when they learn about their scores.

Ideally, we would study deanonymized judge scores both before and after the reform, for example to evaluate how behavior changes for a compatriot judge on the panel compared to non-compatriot judges, or how the same judge behaves under different publication regimes. Unfortunately, it is precisely the anonymization of individual judges' scores that prevents any analyses that require scores to be matched to judge identity before the reform. Therefore, we will mainly focus on judge panel-level statistics such as the aggregate score or the within-panel score dispersion as outcome variables. This implies that we are not able to identify the extent of favoritism by the compatriot judge him-/herself prior to the reform for Non-JGP events. Instead, we will investigate the *aggregate* net bias of a skaters' score when there is a compatriot judge on the panel, which may also include potential favoritism by non-compatriot judges, e.g. due to bloc-voting, as well as strategic counter-exaggerations.

5.2. Estimating effects on score dispersion

In our baseline specification, we estimate the following difference-in-differences model using judge score data at the performance-level:

$$\begin{aligned} \sigma_{isrp} \;=\; & \alpha + \beta_1 \cdot NonJGP_p + \beta_2 \cdot NonJGP_p \times Post_s \\ & + \delta' x_{isrp} + \varphi_s + \varepsilon_{isrp}\,, \end{aligned} \tag{15}$$

where σ_{isrp} is the within-panel standard deviation of scores for performance p by skater i in round r and season s. $NonJGP_p$ is an indicator variable for performances at Non-JGP events. φ_s represents season fixed effects that capture any changes in score statistics over time. The main independent variable of interest is $NonJGP \times Post_s$, which is the interaction of the Non-JGP indicator with an indicator for post-reform events (season 2016-17 onwards). Hence, β_2 is the estimated average effect of the transparency reform on the outcome of interest. We include a number of control variables such as the skater's current ISU world rank.[18] Importantly, we control for a quadratic polynomial of the median score in the panel, as differences in score levels may be linked to higher or lower dispersion across judges, for example due to ceiling effects at the upper score bound.[19] To further test robustness, we also estimate additional specifications with skater fixed effects α_i.

5.3. Estimating effects on nationalistic bias

Identifying biases in performance evaluation is not a straightforward task when scores are anonymized. It is commonly suspected that figure skating judges tend to be positively biased toward skaters with the same nationality, but all we can do without knowledge of individual judges' scores is to compare the aggregate scores for performances by skaters with a compatriot judge on the panel with scores for performances by skaters whose nation is not represented on the panel. Conceptually, this gives us a measure of the *aggregate* bias in the panel that combines behavior by compatriot judges and potential responses by the non-compatriot judges.

The main complication with this comparison is that the presence (or absence) of a compatriot judge on the panel is generally also positively correlated with the skater's skill, because countries with traditionally strong figure skating athletes also tend to be overrepresented in judge panels — judges usually being former competitive skaters themselves. To identify nationalistic bias, we therefore exploit that, from the skater's point of view, the composition of the panel can be regarded as quasi-random. Thus, by including skater fixed effects, we compare scores for the same skater depending on whether he or she performs with a compatriot judge on the panel or not. To hold constant the judge panel and the general performance level of the competitors, we further include skating round fixed effects. The statistical model is then the following:

$$\pi_{irp} = \alpha_i + \beta_1 \cdot Comp_{irp} + \varphi_r \, + \, \delta' x_{irp} \, + \, \varepsilon_{irp} \, , \tag{16}$$

[18] Skaters' world ranks are updated by the ISU after every event, and are computed based on the skater's highest/second highest placements at various sanctioned competitions from the previous two seasons and the current season. Some skaters are not ranked, because they placed too low in previous competitions or because they are new. To account for this, we create an indicator variable for being unranked. Communication No. 1629 (International Skating Union, 2010) provides details regarding rank point distributions.

[19] We use the median rather than the (trimmed) mean score because it is more robust to outliers, which could themselves affect the standard deviation. That said, the correlation is more than 99.8%.

where π_{irp} is the artistic (technical) score a skater i received for performance p in round r, which is calculated as trimmed average score of all judges in the panel. The main regressor of interest here is the indicator variable $Comp_{irp}$, which takes the value 1 if the panel for performance p includes a judge with the same nationality as the performing skater i, and 0 otherwise. Hence, β_1 gives us an estimate of the baseline score gap. α_i and φ_r represent skater and round fixed effects, respectively. In additional specifications, we also control for a vector of other objective skater and performance characteristics x_{irp}, such as skaters' world rank (at the time of performance) and a home event dummy, indicating whether the event took place in a skater's home country, as well as the Base Value, which gives us a performance-level measure that sums up the difficulty of technical elements the skater chose to include in the choreography. Our most stringent specification replaces α_i with skater-season fixed effects α_{is}, thereby accounting for variation in a skater's performance levels throughout the career.[20]

To facilitate interpretation and make scores comparable across a wide range of different events, π_{irp} is normalized across rounds so that its unit is the standard deviation of scores across all performances in round r. This has the additional intuitive appeal that a one-point increase in absolute score is much more impactful for the final rankings when skaters are in a neck-to-neck competition with each other than when their scores are highly dispersed.

After estimating the net degree of nationalistic favoritism in the full sample, we ask whether the transparency reform led to reduction in bias, using the difference-in-differences approach that compares post-reform changes for Non-JGP events relative to JGP events:

$$\begin{aligned}\pi_{irp} = \alpha_i &+ \beta_1 \cdot Comp_{irp} + \beta_2 \cdot Comp \times NonJGP_{irp} \\ &+ \beta_3 \cdot Comp \times Post_{irp} + \beta_4 \cdot Comp \times NonJGP \times Post_{irp} \\ &+ \varphi_r \, + \, \delta' x_{irp} \, + \, \varepsilon_{irp} \, .\end{aligned} \tag{17}$$

Compared to equation 16, we further interact the compatriot performance indicator with an indicator for Non-JGP events ($Comp \times NonJGP_{irp}$), to control for time-invariant differences between the level of favoritism between JGP and Non-JGP events, and with an indicator for post-reform events ($Comp \times Post_{irp}$), to control for common time trends. Crucially, the triple-interaction term $Comp \times NonJGP \times Post_{irp}$ allows us to estimate how the transparency reform affects the compatriot score advantage.

[20] Note that this can heavily affect the implicit weights of observations when identifiying the compatriot score advantage, as for some skaters we observe few or no performances at all with/without a compatriot judges on the panel in a given season.

6. Main Empirical Results

6.1. Effects on average score dispersion

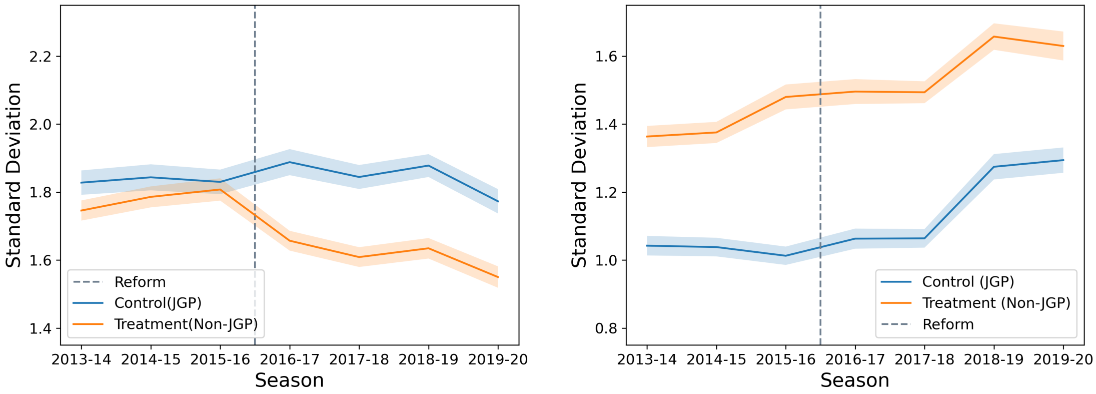
First, we examine whether the transparency reform affected the dispersion of scores across judges for the same performance. Figure 1 plots the average season-by-season within-panel standard deviations of the artistic score and the technical score, respectively, separately for Non-JGP and JGP performances. Reassuringly, the within-panel standard deviations seem to follow parallel trends both in the pre-reform seasons and in the post-reform seasons.[21] But strikingly, there is a sharp drop in the artistic score dispersion for Non-JGP performances after the introduction of the transparency reform in 2016 relative to JGP performances, and this gap persists over time. This provides some first descriptive evidence that judges within a panel award more similar scores for the same performance under transparency than under anonymity. However, we observe no analogous effect for the technical score. While the pre-reform gap between treatment and control performances is much starker, the difference remains more or less constant post-reform. The general increase in the technical score dispersion from season 2018-19 onwards is likely due to a scoring reform that increases the range of possible GOEs that judges can assign from 7 points (-3 to 3 in one-point increments) to 11 points (-5 to 5 in one-point increments).

Table 3 presents the formal difference-in-differences estimates based on regression equation 15. In general, the extent of disagreement among judges follow an inverse-U shaped pattern with regard to the quality of the performance, proxied by the median score — within-panel score dispersion is highest in the middle ranges, whereas scores become more uniform when the performance was either very good or very poor. In contrast, technical score dispersion generally increases with performance quality, because grades are scaled proportionally to the difficulty of the executed elements. Additionally, we observe that the presence of a compatriot judge (with the same nationality as the skater) on the panel is associated with a small but statistically significant increase in score dispersion by around 2%, which is hints at potential score inflation by the compatriot judge due to nationalistic favoritism.

The main coefficient of interest is $Post \times Non-JGP$, which is the indicator for treated events after the transparency reform. The estimates confirm the pattern we observe in Figure 1. Column (1) shows that this coefficient is negative and highly significant for the artistic score, implying that different judges award more similar performance scores in response to the reform. The coefficient of -0.121 ($p = 0.008$) is quantitatively meaningful, corresponding to an effect size of about 21% of a pre-reform standard deviation

[21] To further examine the plausibility of the parallel trend assumption, we plot in the Figure A.2 season-by-season panel standard deviations (as in Figure 1), but with an extended pre-reform period, starting from the 2005-06 season, which is the first season under the current ISU scoring system.

Figure 1: Within-panel standard deviation of scores in JGP (control) and Non-JGP (treated) events



(a) SD of artistic score

(b) SD of technical score

Note: Each point indicates the average panel standard deviation for a season, for JGP (Control, blue) and Non-JGP (Treated, orange) performances. The dashed line indicates the implementation of the transparency reform in 2016; error bars indicate 95% confidence intervals.

(across performances) in panel score dispersion. This decrease in score dispersion that we estimate is also robust to the inclusion of skater fixed effects in column (2), although the coefficient drops slightly to -0.103 ($p = 0.035$). In contrast, there is no effect on the within-performance standard deviation of the technical elements score. While the coefficients are always negative, indicating a decrease in score dispersion, they are quantitatively much smaller and statistically insignificant. This null result stays the same when we only include performances until season 2017-18 in column (5), due to the change in grading scales for the technical score starting from season 2018-19.

We can further break down the score compression effect of the transparency reform into effects across the full distribution of individual performance scores in the panel. To do so, we rank the nine individual scores for any given performance from lowest (1st) to the highest (9th) and calculate their distance to the median score (5th) in the panel. We then use these score distances as dependent variable to estimate the difference-in-differences model on the performance-judge level, i.e., seperately for the lowest score, second-lowest score, and so on. If, for example, a reduction in nationalistic bias was the main driver of lower average score dispersion, we may expect a disproportionate effect at the higher end of the score distribution, which is presumably where compatriot judges are likely to fall into.

Figure 2 plots the estimated coefficients. We can see that after the reform, scores generally becoming more closely packed around the median (for Non-JGP relative to JGP performances). Particularly the extreme scores at either end of the distribution move much

Table 3: Effect of de-anonymized publication on standard deviation of panel scores.

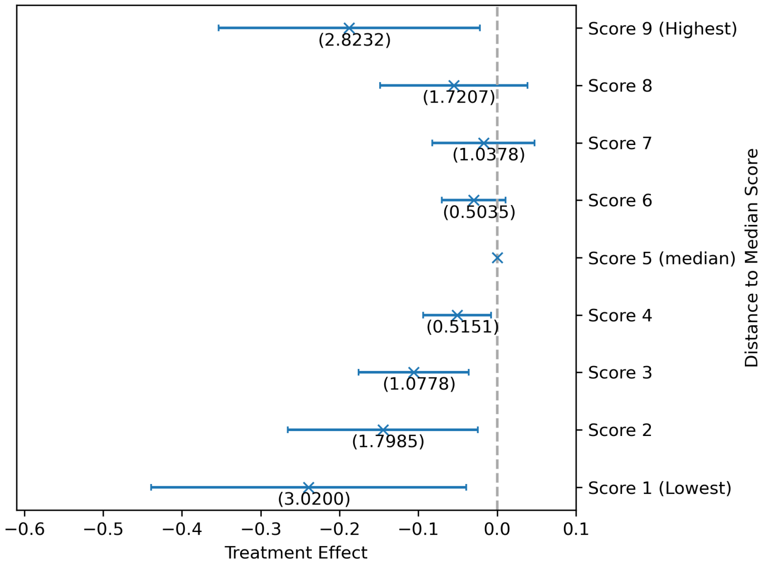
	SD of Artistic Score		SD of Technical Score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	-0.014 (0.041)	-0.033 (0.043)	0.008 (0.020)	-0.018 (0.021)	-0.009 (0.020)
Post × Non-JGP	-0.121*** (0.045)	-0.103** (0.049)	-0.025 (0.028)	-0.034 (0.028)	-0.009 (0.029)
Compatriot	0.038** (0.015)	0.039** (0.015)	0.028*** (0.008)	0.025** (0.010)	0.024* (0.012)
Median score	0.709*** (0.053)	0.694*** (0.097)	0.395*** (0.025)	0.368*** (0.033)	0.272*** (0.032)
Median score squared	-0.099*** (0.007)	-0.107*** (0.011)	0.011*** (0.004)	0.015*** (0.005)	0.025*** (0.005)
Constant	3.479*** (0.111)	3.662*** (0.195)	1.201*** (0.034)	1.209*** (0.038)	1.109*** (0.042)
Skater FEs	—	Yes	—	Yes	Yes
World rank controls	Yes	Yes	Yes	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
JGP mean	1.840	1.840	1.115	1.115	1.044
Observations	16821	16764	16821	16764	12119
R^2	0.141	0.301	0.551	0.615	0.615

Estimates of equation (15), with standard deviation of panel scores as dependent variable. World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. Standard errors clustered at event level (e.g. Olympics 2018). Column (5) excludes the 18-19 and 19-20 seasons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ Pre-reform jgp mean dropping singleton skaters for colns 2, 4 and 5?

closer to the center, implying a reduction in large outliers. Interestingly, the compression pattern is asymmetric, with lower scores on average moving more upwards than higher scores move downwards. The asymmetry is not driven by large outliers and ceiling effects. If the within-panel standard deviation dropped due to a decrease in nationalistic favoritism under transparency, we would expect the opposite, namely an overproportionate effect on positive outliers rather than negative outliers.

Overall, the results in this section show that, in response to the transparency reform, judges award more similar evaluations to their peers' with regard to artistic aspects of a performance, but not with regard to the more objective technical score. This is in line with what our theoretical framework in Section 3 predicts. When facing greater public visibility, reputation concerns can make skaters averse to appearing incompetent or biased when their scores are too out-of-line with fellow judges, in particular in the absence of objective standards against which the public can gauge the accuracy of a judge's scores.

Figure 2: Estimated effect of transparency on distance to the median score, by ranked order



Note: Each point plots the coefficient on Non-JGP×Post, obtained from estimating Equation (15) with the distance of the k-th highest(lowest) score on the panel to the median score as the dependent variable. Controls for discipline×segment, panel median score, panel median score squared and season fixed effects are included. Whiskers indicate 95% confidence intervals (adjusted for clustering at event level); figures in parentheses indicate pre-reform means for Non-JGP (Treat) performances.

As judges cannot communicate with each other and explicitly coordinate their scores, the question thus becomes how the conformity effect comes about. The theoretical framework suggests that higher effort exertion or collective conservatism, i.e., anchoring more towards a common prior, could be potential channels. Another potential channel is that judges curb their idiosyncratic biases toward skaters, with the most prominent source of bias being nationality. In the following, we will explore nationalistic favoritism in judge evaluations and how it was impacted by the transparency reform.

6.2. Effects on nationalistic bias

Next, we look at nationalistic favoritism and how the transparency reform affected the compatriot score advantage, as measured by how much higher the score is for skaters with a compatriot judge on the panel, compared to similar skaters without a compatriot judge on the panel. To make the outcome variable more comparable across rounds, we normalize scores such that one unit corresponds to the standard deviation of scores across skaters within the respective round, and the average performance in each round takes the value

Table 4: Estimated compatriot score advantage in the full sample

	Artistic score (std.)			Technical score (std.)		
	(1)	(2)	(3)	(4)	(5)	(6)
Compatriot	0.066*** (0.010)	0.046*** (0.009)	0.050*** (0.008)	0.044*** (0.014)	0.014** (0.007)	0.020*** (0.007)
Home event	–	0.084*** (0.018)	0.074*** (0.017)	–	0.067*** (0.013)	0.061*** (0.014)
Base value (std.)	–	0.204*** (0.008)	0.133*** (0.007)	–	0.732*** (0.007)	0.706*** (0.008)
World rank controls	–	–	Yes	–	–	Yes
Skater × Season FEs	–	–	Yes	–	–	Yes
Skater FEs	Yes	Yes	–	Yes	Yes	–
Round FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	16764	16764	16589	16764	16764	16589
R^2	0.868	0.891	0.937	0.709	0.911	0.933

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

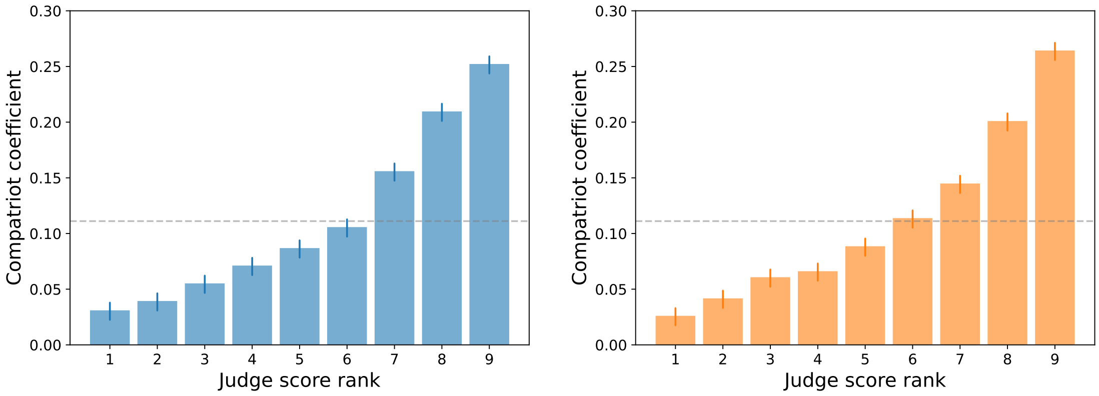
0. This is intuitively appealing, as even a small positive bias in a skater's absolute scores is can result in a sizable relative advantage for the final ranking when all competitors are very close to each other, whereas it would be of little consequence when the competitors' scores are far apart from each other.

6.2.1. Documenting nationalistic bias

We first document a robust and statistically significant score advantage for skaters who have a compatriot judge on the panel and argue that it is likely indicative of nationalistic favoritism in performance evaluation. Table 2 showed that without including controls for ability and other characteristics, skaters with a compatriot judge on the panel receive on average more than 2 points higher raw score in both the artistic and technical domain, compared to their peers without a compatriot on the panel. However, this score gap could be driven by higher performance quality, as judges are more likely recruited from countries that are traditionally strong in figure skating. To control for this, we estimate equation 16, using skater fixed effects to adjust for differences in skater skill, as well as round fixed effects to compare between skaters who compete in the same round and are evaluated by the same panel of judges.

Table 4 columns 1 and 4 show that, once controlling for round and skater fixed effects, the estimated compatriot score advantage in our full sample is about 6.6% of a round-level SDs ($p < 0.001$) for the artistic score and 4.4% for the technical score ($p = 0.002$). When

Figure 3: Distribution of compatriot score rankings towards compatriot performances



(a) Artistic score

(b) Technical score

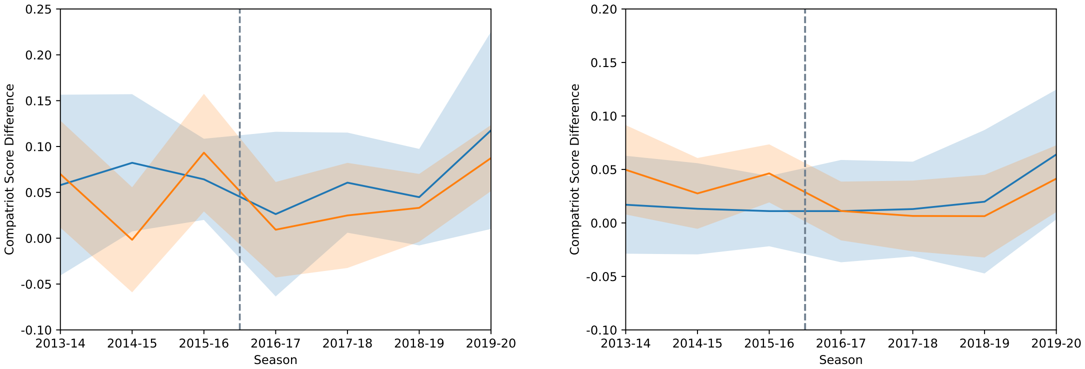
Note: Each bar plots the coefficient from the regression of a binary variable of a particular judge score rank (1 = lowest score, 9 = highest score) against a binary variable indicating whether a judge is a compatriot judge using performance × judge level dataset, with performance fixed effects. Error bars indicate 95% confidence intervals.

adding flexible controls for the skaters' current world rank at the time of competition and the Base Value, which is an objective measure of the performance difficulty, the compatriot effect drops to about 4.6% of a within-round SD for the artistic score and 1.4% for the technical score, but remains highly statistically significant. These estimates stay unchanged when using a stricter specification with skater × season fixed effects that allows us to explain more than 93% of the within-round variation in skaters' performance scores.[22] Our estimates for the aggregate nationalistic bias are quantitatively almost identical to those reported by Zitzewitz (2014).

To further confirm that this residual compatriot score advantage is likely driven by nationalistic bias rather than higher (unobserved) performance quality, we analyze behavior by individual judges on the panel. This restricts our sample to performances under the transparent judging regime, i.e., JGP events and post-reform Non-JGP events. Figure 3 plots the post-reform distribution of judges' score rankings within the panel when they evaluate performances by skaters of the same nationality as themselves. If the compatriot judge was not more likely to award higher scores to a compatriot skater, relative to other judges on the panel, the probability of each score ranking should be 1/9. However, this is clearly not the case. The distribution is heavily left-skewed for both the artistic and technical score, implying that compatriot judge often award unusually generous scores compared to the non-compatriot peers. Indeed, compatriot judges are almost four times

[22] Differences in average scores across rounds in themselves already explain about 85% (71%) of the variation in raw artistic (technical) scores across all skating performances.

Figure 4: Compatriot score advantage for JGP (Control) and Non-JGP (Treated) events



(a) Artistic score

(b) Technical score

Note: Lines indicates the average within-round compatriot score differential by season, separately for JGP (Control) and Non-JGP (Treated) events. We regress the within-round normalized artistic (technical) score on compatriot×season dummies, including round and skater fixed effects, and controlling for home event, within-round normalized base value, squared base value, within-round normalized deductions and squared deductions. Standard errors clustered at event level. (Add legend, check if colors are flipped. Font size a bit too small.) The dashed line indicates the implementation of the transparency reform; error bars are 95% confidence intervals.

as likely to award a score above the panel median than they are to award a below-median score.

Appendix Table A.2 shows that, compared to the non-compatriot judges, a compatriot judge awards a 1.15 points higher overall artistic score and a 1.14 points higher overall GOE score on average for the same performance. Unlike Sandberg (2018), we find no evidence that skaters with a compatriot judge on the panel are evaluated more favorably even by the non-compatriot judges, but there is also no evidence for compensating fairness through strategic counter-exaggeration. Note that judges' evaluations are more impactful for the artistic compared to the technical score, as the letter is determined both by the GOE, awarded by judges, and the objective Base Value, which reflects the difficulty of the performed technical elements.

6.2.2. Effects of higher transparency

Having documented a statistically significant and robust compatriot score advantage that is suggestive of nationalistic bias in performance evaluation, we next turn to the question of whether this score advantage was reduced by the transparency reform, which arguably allowed closer public scrutiny of compatriot judge behavior. As first descriptive evidence, Figure 4 plots the evolution of estimated (within-round) compatriot score differentials

Table 5: Effect of the transparency reform on compatriot score advantage

	Artistic score (std.)		Technical score (std.)	
	(1)	(2)	(3)	(4)
Compatriot	0.070***	0.035*	0.037***	0.032**
	(0.019)	(0.019)	(0.012)	(0.012)
Compatriot × Non-JGP	-0.006	0.018	-0.032*	-0.022
	(0.026)	(0.030)	(0.018)	(0.018)
Compatriot × Post	-0.042*	0.000	-0.035**	-0.024
	(0.024)	(0.023)	(0.015)	(0.018)
Compatriot × Post × Non-JGP	0.040	0.015	0.050**	0.046*
	(0.036)	(0.035)	(0.024)	(0.025)
Home event	0.072***	0.075***	0.065***	0.061***
	(0.017)	(0.017)	(0.013)	(0.014)
Base value (std.)	0.213***	0.133***	0.733***	0.706***
	(0.008)	(0.006)	(0.007)	(0.008)
World rank controls	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16589	16764	16589
R^2	0.885	0.937	0.911	0.933

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

over time, separately for JGP and Non-JGP events. Despite some fluctuations in the order of magnitude that is statistically to be expected, JGP and Non-JGP events do seem to follow roughly similar pre-trends in the three seasons before the reform in our data, thus corroborating our difference-in-difference identification strategy. However, the visual patterns do not show any evidence for a decreasing compatriot score advantage in treated events (Non-JGP) following the transparency reform compared to non-treated events (JGP).

Table 5 presents our formal regression results that implement the estimation strategy described in equation 17. For the artistic score, we find no significant pre-reform difference in the compatriot score advantage between JGP and Non-JGP events, despite individual judges' scores from JGP events already being published openly. For the technical score, we find that the pre-reform bias is slightly stronger for Non-JGP events, if anything. Importantly, we find no evidence for a decrease in the average compatriot bias for treated Non-JGP events relative to JGP events after the reform in 2016. The estimated coefficient of 0.014 for the artistic score is statistically insignificant and goes in the opposite direction.

Based on the coefficients in column 2, the implied estimate for the post-reform compatriot bias at Non-JGP events is positive (0.067) and remains statistically different from zero ($p < 0.001$). For the technical score, the point estimate is also positive (0.046) and marginally statistically significant at the 10% level. Thus, it seems that the transparency reform was unsuccessful in achieving one of its main objective, i.e. to reduce nationalistic favoritism.

The absence of any decrease in the aggregate compatriot score advantage is consistent with our theoretical model from section 3, which predicts that a reduction in individual judges' favoritism may be offset in the aggregate score by conformity motives of other judges. However, due to the anonymity of judges' scores prior to the reform, we cannot, unfortunately, directly investigate how much individual judges' behavior changed due to the transparency reform. Another explanation could be that transparency triggers opposing motives for judges evaluations. For example, public scrutiny and fairness norms would push biased judges to curb their tendencies for favoritism, whereas audiences in the home country as well as national associations that appoint the judges may in fact expect that judges behave in a biased way by skewing scores for their compatriot skaters upwards. For example, Zitzewitz (2006) provides suggestive evidence that national associations tend to appoint judges who are more rather than less biased, which can create perverse incentives for judges to favor compatriot athletes as a signal to their national association.

6.3. The mediating role of public attention

In the theoretical framework from Section 3, we assumed that the channel through which transparency affects judge evaluation behavior is through reputational concerns. This implies that the effects of the transparency reform should be particularly pronounced in highly prestigious events that generate large public attention. To test this, we extend the baseline difference-in-differences model from equation 15 by including interactions of the post-reform Non-JGP indicator with prestige of the competition. We proxy prestige by the average world rank of skater's performing in round r. Thus, we estimate the following regression equation:

$$\begin{aligned}\sigma_{isp} = \alpha &+ \beta_1 \cdot NonJGP_{isp} + \beta_2 \cdot NonJGP \times Post_{isp} \\ &+ \gamma_1 \cdot \mathit{RoundQ} \times NonJGP_{isp} + \gamma_2 \cdot \mathit{RoundQ} \times NonJGP \times Post_{isp} \\ &+ \sum_{k=1}^{2} \delta_k \tilde{\pi}_p^k + \varphi_s + \varepsilon_{isp}\,,\end{aligned} \tag{18}$$

where *RoundQ* is our proxy measure for round quality, computed using the average rank of skaters performing in the the round and, for ease of interpretation, normalized to mean 0 and standard deviation 1 for Non-JGP events. We interact *RoundQ* with the Non-JGP indicator and the post-reform Non-JGP indicator, respectively. The main coefficient

Table 6: Heterogeneous effects on score dispersion by round prestige

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	-0.001	-0.006	0.014	-0.025	-0.027
	(0.038)	(0.041)	(0.021)	(0.025)	(0.024)
Post × Non-JGP	-0.119***	-0.140***	-0.024	-0.032	-0.015
	(0.043)	(0.046)	(0.028)	(0.030)	(0.032)
Round quality × Non-JGP	0.071***	0.063***	0.000	-0.012	-0.016
	(0.015)	(0.017)	(0.012)	(0.014)	(0.015)
Round quality × Non-JGP × Post	-0.080***	-0.087***	0.018	0.008	-0.009
	(0.021)	(0.025)	(0.015)	(0.017)	(0.018)
Compatriot	0.035**	0.037**	0.026***	0.026***	0.028**
	(0.015)	(0.015)	(0.009)	(0.010)	(0.012)
Median score (std)	0.700***	0.620***	0.396***	0.367***	0.268***
	(0.052)	(0.094)	(0.025)	(0.033)	(0.032)
Median score (std) squared	-0.098***	-0.098***	0.011***	0.016***	0.026***
	(0.006)	(0.011)	(0.004)	(0.005)	(0.005)
Skater FEs	—	Yes	—	Yes	Yes
World rank controls	Yes	Yes	Yes	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.142	0.301	0.550	0.615	0.615

Estimates of Equation (15), with standard deviation of panel scores as dependent variable. Standard errors clustered at event level (e.g. Olympics 2018). Column (5) excludes the 18-19 and 19-20 seasons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

of interest here is γ_2, which measures how much the treatment effect of transparency on within-panel score dispersion changes for a one standard deviation increase in round quality. Note that this is not a full triple-differences model. We notably omit the main effects for *RoundQ*, because JGP events, which serve as our control group, are generally less exclusive and prestigious than Non-JGP events; hence, the effect of higher round quality is not comparable between these classes of events, as the complete overlap condition is not fulfilled.

Table 6 presents the results on treatment effect heterogeneity for the within-panel dispersion of both the artistic scores and of the technical scores. We can see from columns (1) and (2) that higher event prestige indeed leads to stronger conformity in judges' artistic scores in response to the transparency reform. A one standard deviation increase in round quality is associated with an additional reduction of the within-panel standard deviation by about 0.08 points post-reform, which corresponds to around two-thirds of the effect at the mean. There is no such pattern with regard to the technical score. Overall, the patterns of heterogeneity we observe are consistent with the hypothesis that the higher degree of conformity, in the form of lower dispersion of (artistic) scores within the panel, is driven by stronger reputation concerns when each judge's score is published openly.

While we found no evidence in Section 6.2 for a decrease in the compatriot score advantage on average following the transparency reform, it is conceivable that publishing individual judges' scores also has differential effects on nationalistic judging depending on how prestigious the event is and how much public attention it thus generates. However, using average world rank of skaters as proxy for public attention as before, we do not find that the aggregate compatriot score advantage in rounds with higher prestige decreases more strongly in response to the reform (see Appendix Table A.4).

7. Investigating Potential Mechanisms

In the previous section, we have found that the transparency reform led to a decrease in the artistic score dispersion within the judge panel. Why is this the case, especially given that judges are not allowed to communicate and coordinate with each other? The theoretical framework suggests several ways through which judges can adjust their scoring behavior to this effect, namely through effort, conservatism, or bias-matching. Which of these mechanisms is at play can lead to diametrically opposed implications for whether the reform improved or worsened the accuracy of overall scores. In this section, we present additional empirical results to further explore these mechanisms. Although we are eventually not able to isolate any specific mechanims, we will explore some of the empirical implications of each mechanism.Lastly, we show that the results are unlikely driven by selection effects due to changes in the composition of judge panels after the reform.

7.1. Consistency as proxy for accuracy

Judges' scores becoming more aligned with each other after the transparency reform could be an indicator for more effort and less noise, but it could also be driven by deliberate attempts to match other judges' scores in an attempt to signal competence in the absence of objectively verifiable yardsticks. Therefore, we explore another potential marker of evaluation accuracy that is arguably less salient as public signal, namely how internally consistent judges are in their evaluations. As described in Section 2, the artistic score (i.e., the program component score) awarded by judges is calculated from subscores for (five) different components of the performance, e.g. skating skills, interpretation of music. Likewise, the technical score is calculated from grades of execution for each technical element (e.g jump, spin) performed by the skater. Using performance-judge-level data, we can thus compute the standard deviation of the artistic (technical) subscores for each judge's evaluation of a given skater performance. A low standard deviation implies a high

Table 7: Effect of transparency on within-judge consistency of scores

	SD of artistic subscores		SD of technical subscores		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.017***	0.012***	0.021	-0.027*	-0.026*
	(0.004)	(0.004)	(0.014)	(0.014)	(0.015)
Post × Non-JGP	-0.016***	-0.017***	0.005	-0.007	0.009
	(0.005)	(0.004)	(0.018)	(0.016)	(0.016)
Median score	0.003	-0.024***	-0.034***	-0.075***	-0.088***
	(0.002)	(0.004)	(0.003)	(0.004)	(0.005)
Median score squared	-0.001***	0.000	-0.007***	-0.005***	-0.006**
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
JGP mean	0.219	0.219	1.034	1.034	1.051
Observations	150458	150458	150431	150431	108675
R^2	0.041	0.090	0.233	0.360	0.342

Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

score consistency, which could be interpreted as confidence in judgement, whereas large variability across subscores could be an indicator for incertitude or arbitrariness.

Using the same difference-in-differences approach as for the main empirical analyses, we test whether the transparency reform lead to a decrease in subscore dispersion at the performance-judge level. Table 7 presents the results. We find that judges indeed become more consistent in their evaluations for artistic score components, but not the technical score components. Columns (1) and (2) show that after the transparency reform, the standard deviation of artistic components drops by 0.016 for Non-JGP performances compared to JGP performances. This effect is statistically significant at the 1% level. However, we find no effect of transparency on within-judge consistency of GOEs awarded for the different technical elements. Hence, our results on within-judge consistency are analogous to the previous findings on the score dispersion across judges in a panel, in that we only find effects for the more subjective and more deliberately assigned artistic scores, but not for the more objective and more spontaneously assigned technical scores. Furthermore, we also find similar heterogeneity patterns as before, with effects of transparency being more pronounced for events that draw higher public attention (see Appendix Table A.7).

However, some ambiguity remains as to whether more consistent scores are indeed an indicator for more accurate performance evaluations. Similarity in subscores could understate the true degree of a performance's variation in the artistic merit across different components. It could even be a mark of laziness, for example if the judge awards the same grade for every artistic subscore — although we note that this happens extremely rarely (0.11% in our sample). Finally, judges may simply use higher consistency as a cheap signaling tool to feign the appearance of competence and thoughtful evaluations (Falk and Zimmermann, 2017).[23]

To argue that the increase in score consistency is likely driven by higher accuracy, we relate it to a number of other proxies for the quality of a judge's evaluations. First, Appendix Table A.8 shows that within-performance, i.e., holding constant the "actual" consistency of the skater's delivery, lower variation across artistic subscores is strongly positively related to how close a judge's score is to the median score in the panel, which is a natural evaluation benchmark for individual judges' scores.[24] This relation appears already in the baseline sample of events with anonymous score reporting, and it is more or less unaffected by the transparency reform (see Appendix Table A.9). Importantly, it is not of purely mechanical nature, as consistency of *artistic* subscores also predicts closeness of the *technical* score to the panel median. Second, higher score consistency is associated with a lesser reliance on the heuristic use of whole numbers — although each artistic component can be rated on a scale from 0.25 to 10.00 in quarter-point increments, almost half (47.96%) of the actual reported subscores have integer values, pointing toward an overuse of integers as cognitive shortcut. We find that a one SD increase in artistic score consistency predicts a 7.6% reduction in the frequency of integer subscores.Third, we use the subsample of JGP events and post-reform Non-JGP events — where individual scores can be linked to judge identity — to show that more experienced judges tend to award scores with higher component consistency (see Appendix Table A.11). This result is partly driven by selection effects rather than pure experience effects, i.e., selective appointment of judges to panels based on prior judging behavior.

Overall, these patterns suggest that within-judge consistency of subscores could plausibly be interpreted as rough proxy for accuracy and confidence in judgement. The transparency reform may thus have partially reduced score dispersion across judges due to genuinely higher effort and evaluation quality.

7.2. Conformity through social learning?

Apart from higher effort toward more accurate evaluations, another mechanism through which scores could become more similar to each other is conservatism, meaning that judges

[23] Note that in their laboratory experiment, response consistency plausibly signals skills because consistent answers across tasks actually corresponds to the correct answers. In our context, the validity of consistency as a signal of skills would depend on how correlated (the audience perceives) the individual score components are.

[24] While the general increase in score conformity across panels may in principle result from implicit coordination on a common prior, the current argumentation hinges on the assumption that when evaluating individual judges within the panel, it is the judges who are closer to the median that have likely been more accurate in their scoring.

award scores that are anchored more towards a presumed consensus score (e.g., a common prior), at the potential loss of signal value from personal assessments. In practice, the question is how judges would be able to form accurate beliefs about a potential consensus score without being able to communicate with each other during performances. One possible answer is that judges can in principle observe and learn about fellow judges' tendencies over time, as the panel remains together throughout a competition round and the aggregate scores are displayed after each performance. The median (average) round includes 12 (16.4) skating performances, which gives judges a reasonable sample to receive feedback about how their own scores compared to the aggregate score. Thus, if transparency induces judges to try to move closer to each other by anticipating and guessing which scores the other judges would report, we should observe that conformity increases the later a performances occurs in a round.

This would be straightforward to test if the order of skating was random. It is, however, not — well-performing skaters tend to skate later in the round. Typically, skaters are placed into starting groups based on their world rank or their placement in the short program, with those who ranked or placed better being assigned to later groups.[25] To generate quasi-exogenous variation, we exploit that the order of performance is randomly determined within the skating groups, and thus plausibly uncorrelated to a skaters' ability, conditional on the group. Grand Prix Series and Final events form an exception, because skating orders are usually determined completely based on previous ranking or placement, so we exclude these events from our analyses in this subsection.

Thus, to test the hypothesis of conformity via social learning over time, we take the difference-in-differences specification from equation 15 and add interactions with skaters' starting number as well as skating group fixed effects:

$$\begin{aligned}\sigma_{irgp} = \alpha &+ \beta_1 \cdot Stnr_{irp} + \beta_2 \cdot Stnr_{irp} \times Post_r \\ &+ \beta_3 \cdot Stnr_{irp} \times NonJGP_r + \beta_4 \cdot Stnr_{irp} \times NonJGP_r \times Post_r \\ &+ \delta' x_{irp} + \varphi_{rg} \; + \; \varepsilon_{igp}\,,\end{aligned} \tag{19}$$

Where $Stnr_{irp}$ is the starting number of skater i in round r, and φ_{rg} represents fixed effects for each skating group g in round r. All else is defined as before. As starting order may have an influence on the generosity of scores (Bruine de Bruin, 2006), we control for the median performance score and its square, as before. The relevant coefficient of interest here is β_4, which estimates whether the conformity effect in response to the transparency reform is stronger or weaker for performances later in a round. If the results in Section

[25] The typical size of a skating group varies. Pooling short- and long-program rounds, starting-order groups tend to be larger for JGP rounds (14), compared to Non-JGP rounds (6.5). This is because JGP short program rounds have completely randomized starting numbers. Draw group sizes are similar for the long program (3.9 for both JGP and Non-JGP rounds).

Table 8: Heterogeneous effects on score dispersion by starting order

	SD of Artistic Score		SD of Technical Score	
	(1)	(2)	(3)	(4)
Starting number	0.001	-0.001	0.000	0.001
	(0.002)	(0.002)	(0.001)	(0.001)
Starting number × Post	-0.003	-0.001	0.001	0.001
	(0.003)	(0.002)	(0.002)	(0.002)
Starting number × Non-JGP	-0.019***	-0.015***	-0.002	-0.000
	(0.006)	(0.005)	(0.003)	(0.004)
Starting number × Non-JGP × Post	0.020**	0.015**	0.005	0.003
	(0.008)	(0.007)	(0.005)	(0.005)
Median score	0.093***	0.082***	0.028***	0.025***
	(0.010)	(0.014)	(0.001)	(0.002)
Median score squared	-0.002***	-0.002***	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.584***	0.913***	0.244***	0.303***
	(0.150)	(0.212)	(0.028)	(0.047)
Skater FEs	—	Yes	—	Yes
Skating group FEs	Yes	Yes	Yes	Yes
Observations	12861	12788	12861	12788
R^2	0.412	0.552	0.739	0.787

Estimates of Equation (19), with standard deviation of panel scores as dependent variable. Standard errors clustered at event level (e.g. Olympics 2018). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.1 are driven by social learning of judges, we should expect β_4 to be negative, indicating a larger decrease in the panel standard deviation for late performances.

Table 8 presents the results for both artistic and technical score. Prior to the reform, the within-panel artistic score dispersion of Non-JGP performances (but not of JGP events) tends to decrease as the round proceeds. Scaling by the average number of skaters in a starting order group, the estimates in column (1) would imply a decrease of 0.078 from the first to the last skater in the group. This could potentially be due to social learning even under anonymous scoring, as judges acquire panel-specific information on scoring with each additional skater, but alternative explanations are also possible — for example, evaluations may become less noisy when judges see more performances that they can use as reference points.

Importantly, we find no evidence of progressively stronger reductions of score conformity when the transparency reform is introduced. Indeed, the estimate on $Non-JGP \times Post \times StNr$ for the artistic score (columns 1 and 2) is positive, and quantitatively similar in absolute value to the estimated coefficient on $Non-JGP \times StNr$. Hence, we find that the tendency to award more similar scores towards later performances in Non-JGP rounds disappears post-reform. Columns (3) and (4) show that the standard deviation of the technical score does not seem to be affected by starting order in any form whatsoever.

We conclude that, for skaters of ex-ante comparable skill, the conformity effect does not seem to vary with starting number, as predicted by social learning. Instead, we find the reversed order effect for the artistic score, which may point to other mechanisms. For example, it is possible that prior to the reform, judges become more deft in their evaluations over time, as they build a reference base of comparable performances against which they can benchmark the current performance. The transparency reform could thus have induced judges to exert greater effort in evaluating earlier performances, so that the observed panel standard deviation becomes more uniform throughout the round.

7.3. Presence of compatriot judges

Anchoring to other judges' scores may not actually require learning and adapting over multiple performances. For example, as discussed in section 3, a decrease in panel score dispersion may be partly driven by judges matching the biases of other judges on the panel. This may be well anticipated ex ante, e.g. in the case of nationalistic favoritism, and therefore do not require any learning over the round. Conformity would create pressure for compatriot judges to adjust their scores downwards, and for the non-compatriot judge to move their score slightly upwards toward the biased judge, so that overall, the score dispersion decreases more for compatriot performances. This mechanism would be consistent with Sandberg (2018), who finds that judges for dressage competitions have a bias towards athletes of the same nationality as other judges on the panel. Alternatively, there might be strategic exaggeration and counter-exaggeration motives among panel judges, for example if judges with fairness concerns want to compensate for favoritism by a compatriot judge on the panel by counter-biasing. Transparency could mitigate such motives, in which case we would observe an even larger drop in the standard deviation of scores of performances with a compatriot judge on the panel. Finally, compatriot performances may simply draw larger public scrutiny, which would lend further support to the notion that reform works by triggering reputation concerns.

Table 9 presents results from fixed effects regressions of within-panel standard deviation on interactions between the treatment status dummies and an indicator for compatriot performances. For all specifications, we include round fixed effects, so that estimates compare skaters of similar skill and facing the same judge panel. With regard to the artistic score, we find some weak evidence to support our hypotheses that scores for compatriot performances become more uniform in response to the transparency reform. The point estimates for the compatriot triple-interaction with Non-JGP and post-reform are negative, indicating an additional conformity effect of transparency in artistic scores of

Table 9: Heterogeneous effects on score dispersion by presence of compatriot judges

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Compatriot	0.019	0.018	0.026^{**}	0.017	0.014
	(0.027)	(0.031)	(0.011)	(0.017)	(0.017)
Compatriot × Non-JGP	0.066^{*}	0.066^{*}	0.010	0.026	0.023
	(0.036)	(0.038)	(0.015)	(0.022)	(0.022)
Compatriot × Post	-0.005	0.029	0.005	0.017	0.007
	(0.034)	(0.040)	(0.014)	(0.020)	(0.021)
Compatriot × Post × Non-JGP	-0.042	-0.087^{*}		-0.022	-0.010
	(0.047)	(0.049)		(0.030)	(0.033)
Median score	0.091^{***}	0.076^{***}	0.026^{***}	0.021^{***}	0.018^{***}
	(0.007)	(0.012)	(0.001)	(0.002)	(0.002)
Median score squared	-0.002^{***}	-0.002^{***}	0.000	0.000^{***}	0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Skater FEs	—	Yes	—	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.315	0.448	0.641	0.693	0.690

Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

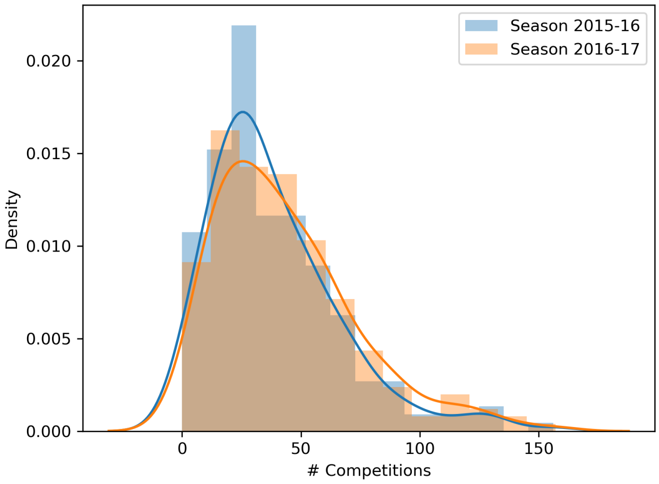
compatriot performances. The coefficient is statistically insignificant ($p = ...$), although it becomes weakly significant when including skater fixed effects ($p = ...$). Quantitatively, it is smaller than the average treatment effect estimates in Table 3, so it compatriot performances alone cannot explain the average score conformity effects, given that a compatriot judge is present for about 67% of Non-JGP performances and is generally even higher for very prestigious events.[26] Overall, we find some weak suggestive evidence that the effects of the transparency reform may be amplified by the presence of compatriot judges on the panel, which could be explained by bias-matching or by larger perceived public scrutiny for these types of performances.

7.4. Composition of judge panels

Finally, we test whether our results on the effect of higher transparency could be explained by changes in the composition of judge panel following the reform, as opposed to changes in the scoring behaviour by individual judges. The process of selecting and appointing judges to a panel is not random and not uniform across events. For JGP events and a small

[26] Recall that the score conformity effect also tends to be stronger per se, as we have shown in Table 6. Additional results controlling for skater's relative rank within the round in Appendix Table A.3 show that the point estimates for the compatriot skater interaction remain similar.

Figure 5: Distribution of Non-JGP judge experience around the transparency reform.



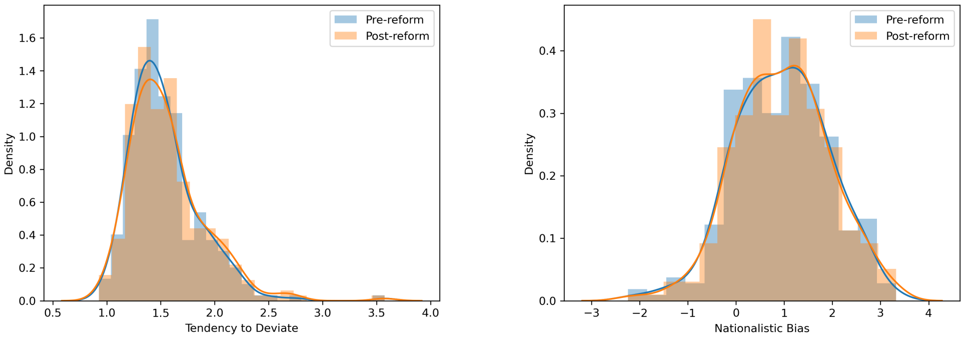
Note: Judge experience in a season is measured as the number of competitions he or she has judged at, from season 2005-06 up to the previous season.

subset of Non-JGP events (the Grand Prix Series), judges are selected by the organizing country.[27] For all other Non-JGP events, judges are selected in a two-step procedure. In the first step, each national skating federation nominates a judge from their country to serve in a particular competition; next, the ISU randomly draws the required number of judges from the pool of proposed candidates. Note that under anonymization, a judge's past scores are also concealed from national skating federations and organizing countries, so that evaluations in JGP competitions were the only objective source of information that federations could use to select judges before the 2016 season.

The observed decrease in score dispersion could be caused by changes in the selection criteria of organizing countries (JGP and GP Series) or national skating federations (all other Non-JGP events) — for instance if under transparent scoring, countries or federations feel compelled to propose judges that are more experienced, less biased, or that have proven more capable in the past. Similarly, potential judges who doubt their own ability may become less willing to serve in panels when they know that their scores will be publicly disclosed. While selection effects can in general be important and meaningful consquences of a transparency reform, we provide several pieces of evidence that speak against these mechanisms.

[27] Selection is subject to the restrictions that judges must come from a pool of qualified individuals ('International Judges') and that no more than one judge from their country is allowed to serve in a given competition. As the Grand Prix Series only feature very few skaters, these events only account for a small fraction of observations in our sample. Our results are robust to dropping these observations.

Figure 6: Distribution of baseline judge-level scoring proxies



(a) Score accuracy proxy by judge

(b) Nationalistic bias proxy by judge

Note: Score accuracy is proxied by a judge's average absolute deviation from the scores given by other judges on the panel. Nationalistic bias is proxied by the difference in the average deviation from other judges' scores for compatriot skater performance relative to non-compatriot performances. Both measures are based on JGP data from the seasons 2005-06 to 2012-13.

First, we check if countries become more likely to select more experienced judges, where we construct a proxy for experience using the number of competitions since the 2005-06 season (the earliest season we can observe) in which a judge has served in a panel. Figure 5 compares histograms of judge experience in the last pre-reform season (2015-16) to the first post-reform season (2016-17). There is no evidence that the distribution changes significantly from pre-reform to post-reform (p-value of Kolmogorov-Smirnov test = 0.1397), and inspecting the distribution of judge experience across all seasons in our sample (Appendix Figure A.4) does not reveal any major upward shifts either.

Next, we investigate whether judges selected after the transparency reform differ in their revealed baseline scoring behavior. As the pre-reform Non-JGP results are anonymized, we use data from JGP events over the 2005-06 season to the 2012-13 season, where scores were transparent even before the reform. This allows us to construct individual-level judging measures for about 80% of the judges in our sample. As proxy measure of a judge's scoring accuracy, we calculate the average absolute deviation of a judge's scores from scores by the fellow judges on the panel (see, e.g., Heiniger and Mercier, 2021). As proxy measure of a judge's impartiality regarding nationalistic judging, we calculate the average deviation of a judge's scores from other scores in the panel for performances where the skater is a compatriot, relative to the average deviation in performances where the skater is not a compatriot. Figure 6 shows that, comparing the last pre-reform to the first post-reform season, there do not appear to be significant shifts in the distribution of judges, neither based on baseline score accurace nor on baseline bias.[28]

Finally, we directly examine potential opting-out of Non-JGP events after the introduction of transparent scoring by following the "careers" of judges who have served in Non-JGP event prior to the reform — which also includes judges who are not represented in the previous analysis. Appendix Tables A.13 and A.14 show that there is no significant extensive or intensive margin decrease in judges' propensity to serve in Non-JGP event following the transparency reform. Thus, we find little overall evidence that the conformity effect induced by the transparency reform could be plausibly driven by selection effects rather than effects on individual judging behavior.

8. Conclusion

In this paper, we studied the effect of transparency on performance evaluation in committees in a high-stakes, professional context. Specifically, we evaluated a reform implemented in the sport of figure skating that increased the visibility of judges' decisions. Prior to the reform, judges' scores were published anonymously, thus shielding the judge from public censure or supervision. While this prevents judges from being swayed by public opinion and coerced into collusion by their fellow judges, this opacity also made it was relatively easy for judges to engage in nationalistic favoritism, so that, following accusations of nationalistic judging in the 2014 Sochi Olympics, the ISU de-anonymized result publication for all events.

To illustrate how increased visibility might impact judges' scoring behavior, we proposed a theoretical framework à la Morris and Shin (2002) with potentially biased and conformist judges, in which the transparency reform enters as an increase in conformist concerns. In line with the predictions of the model, we find that the within-performance score dispersion for artistic scores decreases sharply post-reform, indicating that judges tend to award more similar scores. In further support of a conformity-based explanation, we also see that this effect is stronger in settings with greater public attention, where judges might feel higher pressure to conform. Lastly, we find that skaters are scored higher when they have a compatriot judge on the panel, and that this compatriot advantage does not decrease post-reform. This is, at first glance, perhaps surprising, given that the reform was implemented precisely to address such concerns. However, this finding is compatible with our model's predictions, and highlights the limited impact that greater transparency can have on aggregate biases in committee decisions.

Though the sharp increase in scoring similarity is in line with previous research in different contexts, the inability of judges to communicate with each other in our setting rules

[28] For histograms of judge scoring behaviour across all seasons in our estimation sample, see Appendix Figures A.5 and A.6.

out informational exchange or persuasion as mechanisms driving the conformity effect we see. Similarly, we do not find any evidence of social learning in our setting. Our model instead suggests two potential sources for this result— increased effort leading to higher signal precision, or herding on a common prior— with largely different welfare consequences. The former leads to less arbitrary and random scoring, whereas the latter has the opposite effect, and could over time lead to a more entrenched system where performances by rookie skaters are insufficiently rewarded. We ultimately cannot distinguish between these channels with our data, and leave this as a potential avenue to explore in future research.

In general, transparency, by activating social image concerns, is a powerful tool that can be used to align individual behavior with public norms and expectations. Whether this can be successfully utilized to achieve desirable committee outcomes, however, likely depends on a variety of factors. These include, among others, the prevailing norms in the society, the degree of subjectivity of the decision, and the composition of the committee, which influence the quality of decisions made under transparency. Thus, policy makers should carefully consider the context when implementing transparency policies. However, one advantage of higher transparency is hardly disputable: it generates publicly available data for third parties like journalists and researchers and thereby potentially long-term value.

References

Asch, Solomon E. 1951. "Effects of group pressure upon the modification and distortion of judgment." In *Groups, leadership and men; research in human relations.* , ed. H. Guetzkow, 177–190. Pittsburgh:Carnegie Press.

Bagues, Manuel F., and Berta Esteve-Volart. 2010. "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment." *Review of Economic Studies*, 77(4): 1301–1328.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review*, 107(4): 1207–1238.

Baker, Michael, Yosh Halberstam, Kory Kroft, Alexandre Mas, and Derek Messacar. 2019. "Pay Transparency and the Gender Gap: Working Paper."

Bar-Isaac, Heski. 2012. "Transparency, Career Concerns, and Incentives for Acquiring Expertise." *The B.E. Journal of Theoretical Economics*, 12(1).

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015. "Unintended Effects of Anonymous Résumés." *American Economic Journal: Applied Economics*, 7(3): 1–27.

Benesch, Christine, Monika Bütler, and Katharina E. Hofer. 2018. "Transparency in parliamentary voting." *Journal of Public Economics*, 163: 60–76.

Bertrand, Marianne, Sandra E. Black, Sissel Jensen, and Adriana Lleras-Muney. 2018. "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway." *Review of Economic Studies*, 86(1): 191–239.

Böheim, René, Mario Lackner, and Wilhelm Wagner. 2020. "Raising the Bar: Causal Evidence on Gender Differences in Risk-Taking from a Natural Experiment." *IZA Discussion Paper No. 12946.*

Bruine de Bruin, Wändi. 2006. "Save the last dance II: unwanted serial position effects in figure skating judgments." *Acta psychologica*, 123(3): 299–311.

Bursztyn, Leonardo, and Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *Quarterly Journal of Economics*, 130(3): 1329–1367.

Campbell, Bryan, and John W. Galbraith. 1996. "Nonparametric Tests of the Unbiasedness of Olympic Figure-Skating Judgments." *The Statistician*, 45(4): 521–526.

Colombo, Luca, and Gianluca Femminis. 2008. "The social value of public information with costly information acquisition." *Economics Letters*, 100(2): 196–199.

Darby, Michael R., and Edi Karni. 1973. "Free Competition and the Optimal Amount of Fraud." *Journal of Law and Economics*, 16(1): 67–88.

Dohmen, Thomas J. 2008*a*. "Do professionals choke under pressure?" *Journal of Economic Behavior & Organization*, 65(3-4): 636–653.

Dohmen, Thomas J. 2008*b*. "The Influence of Social Forces: Evidence From The Behavior of Football Referees." *Economic Inquiry*, 46(3): 411–424.

Dulleck, Uwe, and Rudolf Kerschbamer. 2006. "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods." *Journal of Economic Literature*, 44(1): 5–42.

Falk, Armin, and Florian Zimmermann. 2017. "Consistency as a Signal of Skills." *Management Science*, 63(7): 2197–2210.

Fehrler, Sebastian, and Moritz Janas. 2021. "Delegation to a Group." *Management Science*, 67(6): 3714–3743.

Fehrler, Sebastian, and Niall Hughes. 2018. “How Transparency Kills Information Aggregation: Theory and Experiment.” *American Economic Journal: Microeconomics*, 10(1): 181–209.

Fernando, A. Nilesh, and Siddharth Eapen George. 2021. “Debiasing Discriminators: Evidence from the Introduction of Neutral Umpires.” *Working paper.*

Garicano, Luis, Ignacio Palacios-Huerta, and Canice Prendergast. 2005. “Favoritism Under Social Pressure.” *Review of Economics and Statistics*, 87(2): 208–216.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. “Social pressure and voter turnout: Evidence from a large-scale field experiment.” *American Political Science Review*, 102(1): 33–48.

Gersbach, Hans, and Volker Hahn. 2012. “Information acquisition and transparency in committees.” *International Journal of Game Theory*, 41(2): 427–453.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. “Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach.” *Quarterly Journal of Economics*, 133(2): 801–870.

Heiniger, Sandro, and Hugues Mercier. 2021. “Judging the judges: evaluating the accuracy and national bias of international gymnastics judges.” *Journal of Quantitative Analysis in Sports*, 0(0).

Huber, Jürgen, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L. Smith. 2022. “Nobel and novice: Author prominence affects peer review.” *Proceedings of the National Academy of Sciences of the United States of America*, 119(41): e2205779119.

International Skating Union. 2010. “Communication No. 1629.”

Jiang, Lingqing. 2020. “Splash with a teammate: Peer effects in high-stakes tournaments.” *Journal of Economic Behavior & Organization*, 171: 165–188.

Kahneman, Daniel, Olivier Sobony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgement.* New York:Little, Brown Spark.

Kim, Jerry W., and Brayden G. King. 2014. “Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring.” *Management Science*, 60(11): 2619–2644.

Krause, Annabelle, Ulf Rinne, and Klaus F. Zimmermann. 2012. “Anonymous job applications of fresh Ph.D. economists.” *Economics Letters*, 117(2): 441–444.

Lee, Jungmin. 2008. “Outlier Aversion in Subjective Evaluation: Evidence From World Figure Skating Championships.” *Journal of Sports Economics*, 9(2): 141–159.

Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review*, 97(1): 150–168.

Lichter, Andreas, Nico Pestel, and Eric Sommer. 2017. "Productivity effects of air pollution: Evidence from professional soccer." *Labour Economics*, 48: 54–66.

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics*, 9(2): 60–92.

Li, Hao, Sherwin Rosen, and Wing Suen. 2001. "Conflicts and Common Interests in Committees." *American Economic Review*, 91(5): 1478–1497.

Litman, Cheryl, and Thomas Stratmann. 2018. "Judging on thin ice: the effects of group membership on evaluation." *Oxford Economic Papers*, 70(3): 763–783.

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. "How social influence can undermine the wisdom of crowd effect." *Proceedings of the National Academy of Sciences of the United States of America*, 108(22): 9020–9025.

Maida, Agata, and Andrea Weber. 2019. "Female leadership and gender gap within firms: Evidence from an italian board reform." *ILR Review*, 0019793920961995.

Mas, Alexandre. 2017. "Does Transparency Lead to Pay Compression?" *Journal of Political Economy*, 125(5): 1683–1721.

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review*, 99(1): 112–145.

Mattozzi, Andrea, and Marco Y. Nakaguma. 2019. "Public versus Secret Voting in Committees." *Working paper.*

Meade, Ellen E., and David Stasavage. 2008. "Publicity of Debate and the Incentive to Dissent: Evidence from the US Federal Reserve." *Economic Journal*, 118(528): 695–717.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science*, 159(3810): 56–63.

Morris, Stephen, and Hyun Song Shin. 2002. "Social Value of Public Information." *American Economic Review*, 92(5): 1521–1534.

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review*, 101(4): 1410–1435.

Pope, Devin G., and Maurice E. Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review*, 101(1): 129–157.

Pope, Devin G., Joseph Price, and Justin Wolfers. 2018. "Awareness reduces racial bias." *Management Science*, 64(11): 4988–4995.

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review*, 95(3): 862–877.

Prendergast, Canice. 1993. "A Theory of "Yes Men"." *American Economic Review*, 83(4): 757–770.

Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination Among NBA Referees." *Quarterly Journal of Economics*, 125(4): 1859–1887.

Rausser, Gordon C., Leo K. Simon, and Jinhua Zhao. 2015. "Rational exaggeration and counter-exaggeration in information aggregation games." *Economic Theory*, 59(1): 109–146.

Rosar, Frank. 2015. "Continuous decisions by a committee: Median versus average mechanisms." *Journal of Economic Theory*, 159: 15–65.

Sandberg, Anna. 2018. "Competing Identities: A Field Study of In-group Bias Among Professional Evaluators." *Economic Journal*, 128(613): 2131–2159.

Stasavage, David. 2007. "Polarization and Publicity: Rethinking the Benefits of Deliberative Democracy." *Journal of Politics*, 69(1): 59–72.

Suurmond, Guido, Otto H. Swank, and Bauke Visser. 2004. "On the bad reputation of reputational concerns." *Journal of Public Economics*, 88(12): 2817–2838.

Swank, Job, Otto H. Swank, and Bauke Visser. 2008. "How Committees of Experts Interact with the outside World: Some Theory, and Evidence from the Fomc." *Journal of the European Economic Association*, 6(2-3): 478–486.

Swank, Otto H., and Bauke Visser. 2021. "Committees as Active Audiences: Reputation Concerns and Information Acquisition." *Working paper.*

Visser, Bauke, and Otto H. Swank. 2007. "On Committees of Experts." *Quarterly Journal of Economics*, 112(1): 337–372.

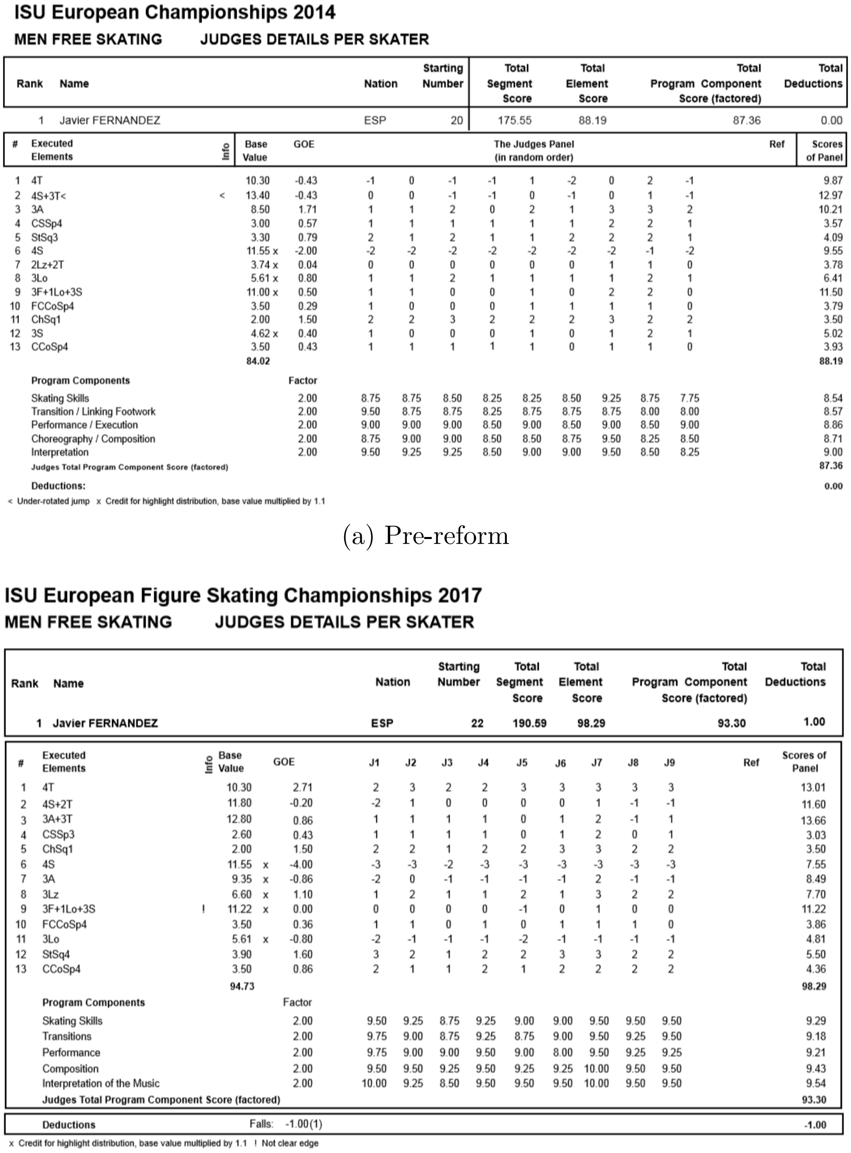
Zitzewitz, Eric. 2006. "Nationalism in Winter Sports Judging and Its Lessons for Organizational Decision Making." *Journal of Economics & Management Strategy*, 15(1): 67–99.

Zitzewitz, Eric. 2014. "Does Transparency Reduce Favoritism and Corruption? Evidence From the Reform of Figure Skating Judging." *Journal of Sports Economics*, 15(1): 3–30.

A. Appendix

A.1. Supplementary figures

Figure A.1: Online publication of results for Non-JGP (Treat) events pre- and post-reform.



ISU European Championships 2014

MEN FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	Javier FERNANDEZ	ESP	20	175.55	88.19	87.36	0.00

#	Executed Elements	Info	Base Value	GOE	The Judges Panel (in random order)									Ref	Scores of Panel
1	4T		10.30	-0.43	-1	0	-1	-1	1	-2	0	2	-1		9.87
2	4S+3T<	<	13.40	-0.43	0	0	-1	-1	0	-1	0	1	-1		12.97
3	3A		8.50	1.71	1	1	2	0	2	1	3	3	2		10.21
4	CSSp4		3.00	0.57	1	1	1	1	1	1	2	2	1		3.57
5	StSq3		3.30	0.79	2	1	2	1	1	2	2	2	1		4.09
6	4S		11.55 x	-2.00	-2	-2	-2	-2	-2	-2	-2	-1	-2		9.55
7	2Lz+2T		3.74 x	0.04	0	0	0	0	0	0	1	1	0		3.78
8	3Lo		5.61 x	0.80	1	1	2	1	1	1	1	2	1		6.41
9	3F+1Lo+3S		11.00 x	0.50	1	1	0	0	1	0	2	2	0		11.50
10	FCCoSp4		3.50	0.29	1	0	0	0	1	1	1	1	0		3.79
11	ChSq1		2.00	1.50	2	2	3	2	2	2	3	2	2		3.50
12	3S		4.62 x	0.40	1	0	0	0	1	0	1	2	1		5.02
13	CCoSp4		3.50	0.43	1	1	1	1	1	0	1	1	0		3.93
			84.02												**88.19**
	Program Components			Factor											
	Skating Skills			2.00	8.75	8.75	8.50	8.25	8.25	8.50	9.25	8.75	7.75		8.54
	Transition / Linking Footwork			2.00	9.50	8.75	8.75	8.25	8.75	8.75	8.75	8.00	8.00		8.57
	Performance / Execution			2.00	9.00	9.00	9.00	8.50	9.00	8.50	9.00	8.50	9.00		8.86
	Choreography / Composition			2.00	8.75	9.00	9.00	8.50	8.50	8.75	9.50	8.25	8.50		8.71
	Interpretation			2.00	9.50	9.25	9.25	8.50	9.00	9.00	9.50	8.50	8.25		9.00
	Judges Total Program Component Score (factored)														**87.36**
	Deductions:														**0.00**

< Under-rotated jump x Credit for highlight distribution, base value multiplied by 1.1

(a) Pre-reform

ISU European Figure Skating Championships 2017

MEN FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score	Total Element Score	Total Program Component Score (factored)	Total Deductions
1	**Javier FERNANDEZ**	**ESP**	**22**	**190.59**	**98.29**	**93.30**	**1.00**

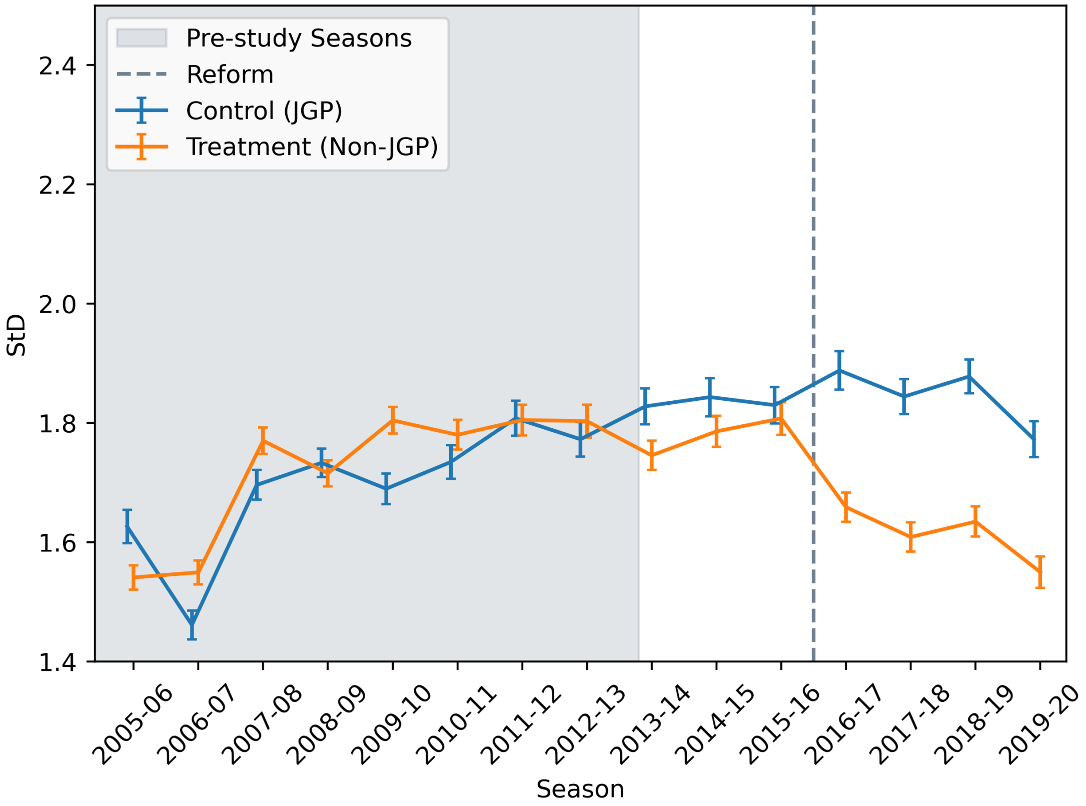
#	Executed Elements	Info	Base Value		GOE	J1	J2	J3	J4	J5	J6	J7	J8	J9	Ref	Scores of Panel
1	4T		10.30		2.71	2	3	2	2	3	3	3	3	3		13.01
2	4S+2T		11.80		-0.20	-2	1	0	0	0	0	1	-1	-1		11.60
3	3A+3T		12.80		0.86	1	1	1	1	0	1	2	-1	1		13.66
4	CSSp3		2.60		0.43	1	1	1	1	0	1	2	0	1		3.03
5	ChSq1		2.00		1.50	2	2	1	2	2	3	3	2	2		3.50
6	4S		11.55	x	-4.00	-3	-3	-2	-3	-3	-3	-3	-3	-3		7.55
7	3A		9.35	x	-0.86	-2	0	-1	-1	-1	-1	2	-1	-1		8.49
8	3Lz		6.60	x	1.10	1	2	1	1	2	1	3	2	2		7.70
9	3F+1Lo+3S	!	11.22	x	0.00	0	0	0	0	-1	0	1	0	0		11.22
10	FCCoSp4		3.50		0.36	1	1	0	1	0	1	1	1	0		3.86
11	3Lo		5.61	x	-0.80	-2	-1	-1	-1	-2	-1	-1	-1	-1		4.81
12	StSq4		3.90		1.60	3	2	1	2	2	3	3	2	2		5.50
13	CCoSp4		3.50		0.86	2	1	1	2	1	2	2	2	2		4.36
			94.73													**98.29**
	Program Components				Factor											
	Skating Skills				2.00	9.50	9.25	8.75	9.25	9.00	9.00	9.50	9.50	9.50		9.29
	Transitions				2.00	9.75	9.00	8.75	9.25	8.75	9.00	9.50	9.25	9.50		9.18
	Performance				2.00	9.75	9.00	9.00	9.50	9.00	8.00	9.50	9.25	9.25		9.21
	Composition				2.00	9.50	9.50	9.25	9.50	9.25	9.25	10.00	9.50	9.50		9.43
	Interpretation of the Music				2.00	10.00	9.25	8.50	9.50	9.50	9.50	10.00	9.50	9.50		9.54
	Judges Total Program Component Score (factored)															**93.30**
	Deductions		Falls:		-1.00(1)											**-1.00**

x Credit for highlight distribution, base value multiplied by 1.1 ! Not clear edge

(b) Post-reform

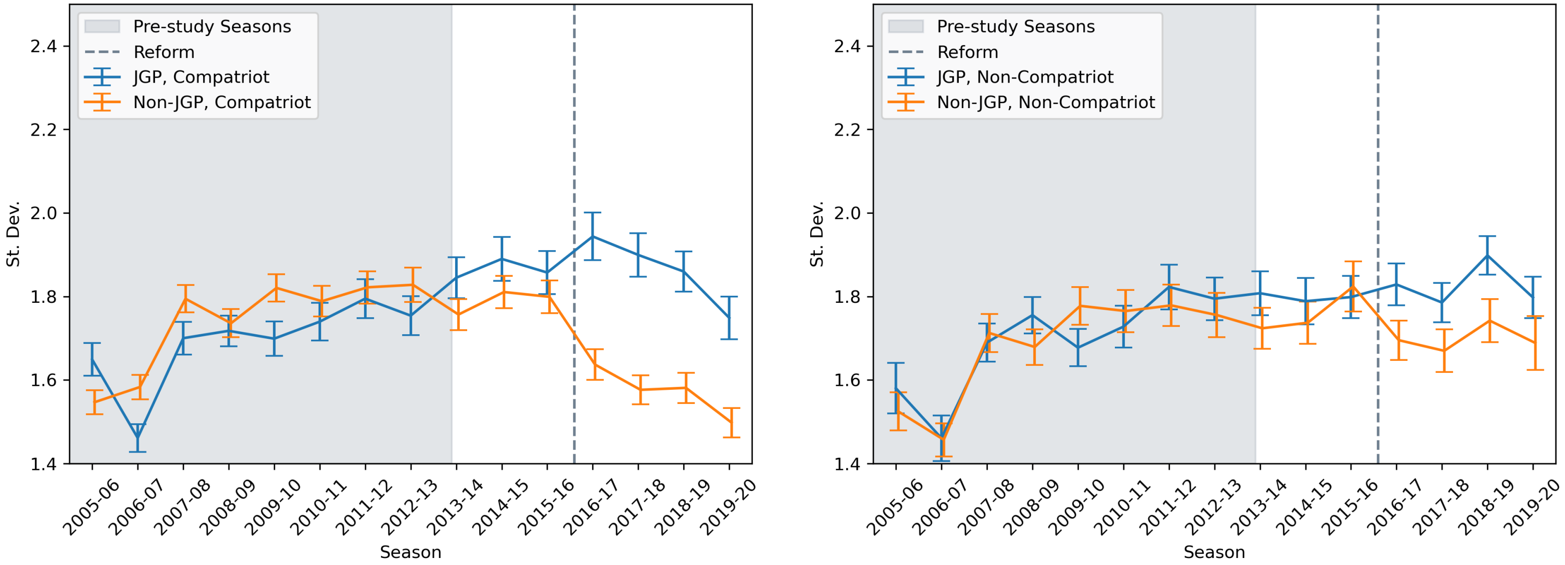
Note: Notice that the order of panel judges is not revealed in panel (a), while it is revealed in panel (b). This order can be linked back to the individual judges on the panel.

Figure A.2: Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2005-06 to 2019-20



Note: Each orange(blue) point plots the average panel standard devation for treatment(control) performances in a season, over the seasons 2005-06 to 2019-20. The dashed line indicates implementation of the transparency reform, from the 2016-17 season onwards.

Figure A.3: Standard deviation of panel scores for JGP (Control) and Non-JGP (Treat) events, from seasons 2013-14 to 2019-20, split by presence of compatriot judge on panel.

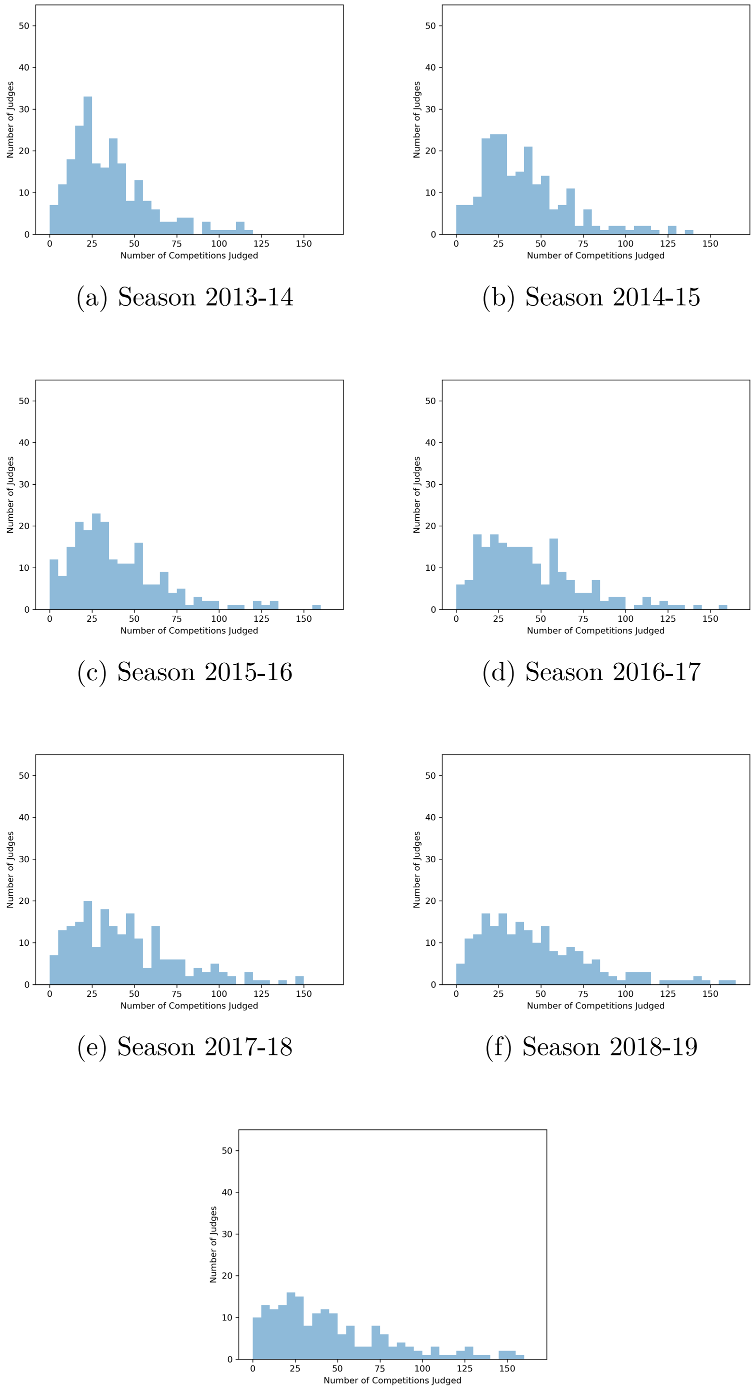


(a) Compatriot

(b) Non Compatriot

Note: Each orange(blue) point plots the average panel standard deviation for treatment(control) performances in a season, over the seasons 2005-06 to 2019-20. The dashed line indicates implementation of the transparency reform, from the 2016-17 season onwards.

Figure A.4: Distributions of Non-JGP (Treat) judge experience by season, from seasons 2013-14 to 2019-20.



(a) Season 2013-14

(b) Season 2014-15

(c) Season 2015-16

(d) Season 2016-17

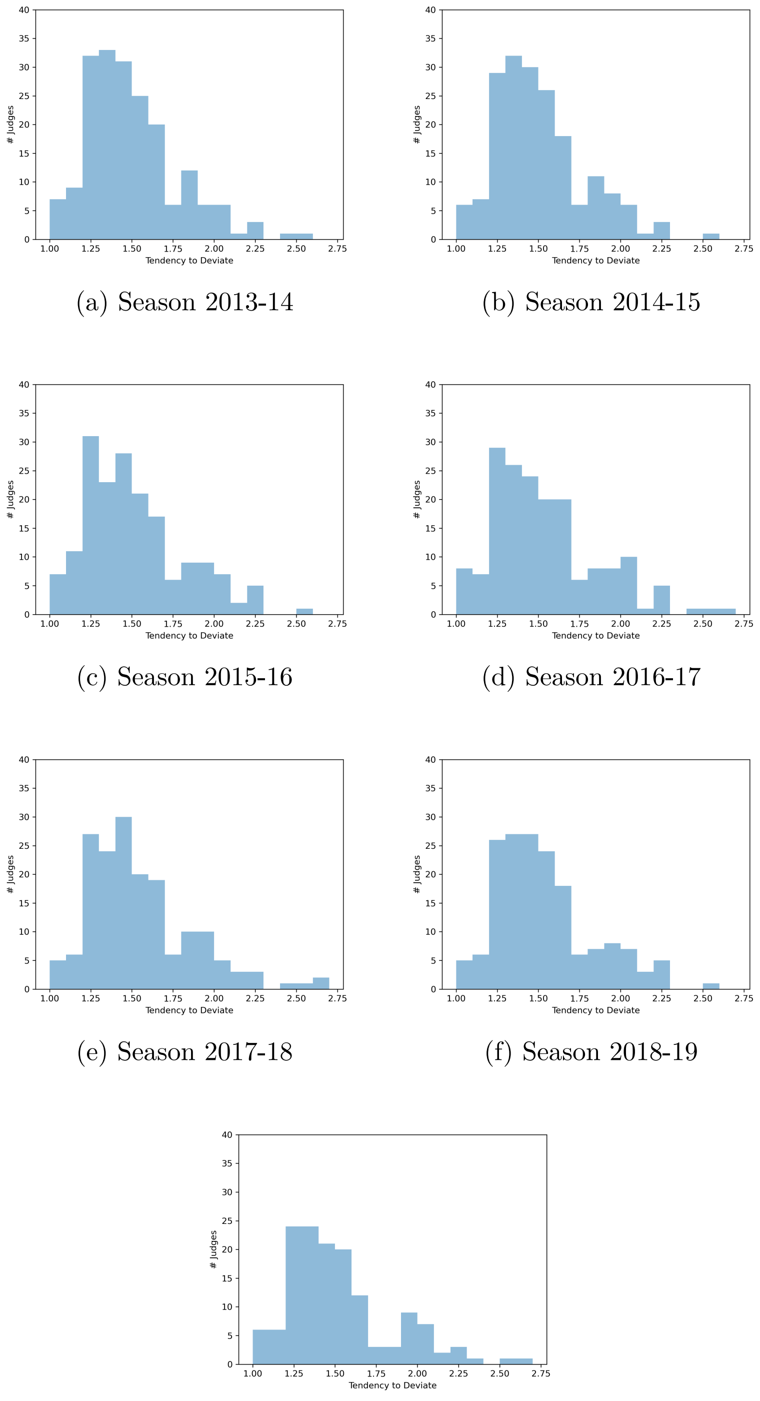
(e) Season 2017-18

(f) Season 2018-19

(g) Season 2019-20

Note: Judge experience in a season is computed as the number of competitions he/she has judged at up until that season.

Figure A.5: Distributions of Non-JGP score accuracy by season, from seasons 2013-14 to 2019-20.



(a) Season 2013-14

(b) Season 2014-15

(c) Season 2015-16

(d) Season 2016-17

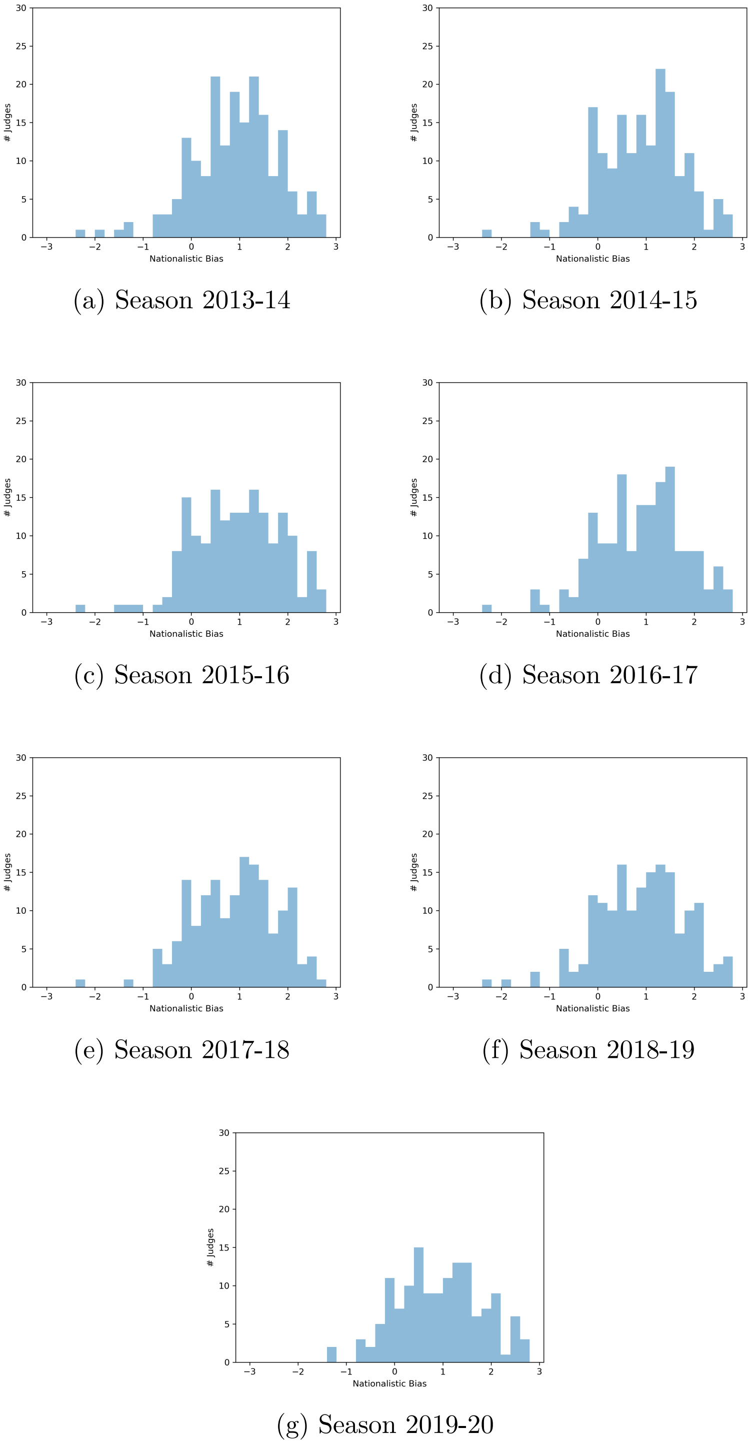
(e) Season 2017-18

(f) Season 2018-19

(g) Season 2019-20

Note: For each judge, his/her measure of deviation is the average deviation of all performances where he/she has judged in, where his/her deviation in a performance is calculated as the absolute value of his score from that of the leave-one-out panel mean.

Figure A.6: Distributions of Non-JGP nationalistic bias by season, from seasons 2013-14 to 2019-20.



(a) Season 2013-14

(b) Season 2014-15

(c) Season 2015-16

(d) Season 2016-17

(e) Season 2017-18

(f) Season 2018-19

(g) Season 2019-20

Note: For each judge, his/her measure of (nationalistic) impartiality is the average deviation from the leave-one-out panel mean when the skater is compatriot, minus the the average deviation from the leave-one-out panel mean when the skater is non-compatriot.

A.2. Supplementary tables

Table A.1: Estimated compatriot score advantage in the full sample

	Artistic score		Technical score	
	(1)	(2)	(3)	(4)
Compatriot	0.052^{***} (0.009)	0.052^{***} (0.009)	0.033^{**} (0.014)	0.033^{**} (0.014)
Home event	0.083^{***} (0.018)	0.083^{***} (0.018)	0.115^{***} (0.024)	0.109^{***} (0.024)
World rank controls	—	Yes	—	Yes
Skater × Season FEs	Yes	Yes	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes
Observations	16589	16589	16589	16589
R^2	0.931	0.931	0.794	0.795

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.2: Compatriot score advantage

	Artistic score		Technical score	
	(1)	(2)	(3)	(4)
Compatriot Judge	1.156***	1.156***	1.142***	1.142***
	(0.030)	(0.031)	(0.043)	(0.044)
Compatriot	0.068	0.055*	0.063	0.052
	(0.042)	(0.031)	(0.066)	(0.078)
Home event	0.472***	0.488***	0.504***	0.425***
	(0.070)	(0.056)	(0.112)	(0.102)
Base Value	0.135***	0.089***	1.137***	1.133***
	(0.004)	(0.003)	(0.010)	(0.011)
Controls for current world rank	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
Judge × Round FEs	Yes	Yes	Yes	Yes
Observations	109296	109296	109296	109296
R^2	0.936	0.950	0.977	0.981

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: Heterogeneous effects within rounds

	SD of artistic score		SD of technical score		
	(1)	(2)	(3)	(4)	(5)
Compatriot	0.019	0.019	0.018	0.018	0.015
	(0.028)	(0.031)	(0.011)	(0.017)	(0.017)
Compatriot × Non-JGP	0.063*	0.061	0.022	0.022	0.020
	(0.036)	(0.038)	(0.022)	(0.022)	(0.022)
Compatriot × Post	-0.004	0.028	0.014	0.014	0.004
	(0.034)	(0.039)	(0.016)	(0.020)	(0.021)
Compatriot × Post × Non-JGP	-0.038	-0.080	-0.009	-0.014	-0.003
	(0.046)	(0.048)	(0.030)	(0.030)	(0.034)
Relative rank	0.046	-0.047	0.091***	-0.023	-0.054
	(0.055)	(0.055)	(0.030)	(0.036)	(0.036)
Relative rank × Non-JGP	0.031	0.143**	-0.027	0.075*	0.069
	(0.068)	(0.069)	(0.040)	(0.044)	(0.044)
Relative rank × Post	-0.016	0.050	0.029	0.068	0.067
	(0.065)	(0.072)	(0.035)	(0.048)	(0.057)
Relative rank × Non-JGP × Post	-0.039	-0.139	-0.158**	-0.143**	-0.120
	(0.084)	(0.088)	(0.062)	(0.063)	(0.081)
Median score	0.092***	0.075***	0.024***	0.021***	0.018***
	(0.007)	(0.012)	(0.001)	(0.002)	(0.002)
Median score squared	-0.002***	-0.002***	0.000*	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Skater FEs	—	Yes	—	Yes	Yes
Round FEs	Yes	Yes	Yes	Yes	Yes
Observations	16821	16764	16821	16764	12119
R^2	0.315	0.448	0.643	0.694	0.690

Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Heterogeneous effects on compatriot score advantage

	Artistic score (std.)		Technical score (std.)	
	(1)	(2)	(3)	(4)
Compatriot	0.053^{***}	0.034^{*}	0.035^{***}	0.031^{**}
	(0.020)	(0.018)	(0.012)	(0.013)
Comp. × Non-JGP	0.010	0.025	-0.031^{*}	-0.020
	(0.031)	(0.032)	(0.018)	(0.018)
Comp. × Post	-0.034	0.002	-0.034^{**}	-0.024
	(0.025)	(0.023)	(0.015)	(0.018)
Comp. × Post × Non-JGP	0.030	0.010	0.051^{**}	0.047^{*}
	(0.039)	(0.037)	(0.025)	(0.026)
Comp. × Round quality × Non-JGP	0.009	0.029	-0.004	0.008
	(0.018)	(0.021)	(0.012)	(0.012)
Comp. × Round qual. × Non-JGP × Post	0.002	-0.020	0.011	0.003
	(0.022)	(0.024)	(0.016)	(0.015)
Home event	0.084^{***}	0.074^{***}	0.066^{***}	0.061^{***}
	(0.018)	(0.017)	(0.013)	(0.014)
Base value (std.)	0.206^{***}	0.136^{***}	0.733^{***}	0.707^{***}
	(0.008)	(0.007)	(0.007)	(0.008)
Controls for current world rank	Yes	Yes	Yes	Yes
Skater × Season FEs	–	Yes	–	Yes
Skater FEs	Yes	–	Yes	–
Round FEs	Yes	Yes	Yes	Yes
Observations	16764	16589	16764	16589
R^2	0.891	0.937	0.911	0.933

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: Effect of de-anonymized publication on variance of panel scores

	Artistic score	Technical score
	(1)	(2)
Compatriot	0.131	0.039
	(0.138)	(0.051)
Compatriot × Non-JGP	0.208	0.073
	(0.168)	(0.074)
Compatriot × Post	0.140	0.100
	(0.187)	(0.071)
Compatriot × Post × Non-JGP	-0.455**	-0.007
	(0.229)	(0.137)
Home event	-0.113*	-0.008
	(0.065)	(0.060)
Controls for current world rank	Yes	Yes
Skater FEs	Yes	Yes
Round FEs	Yes	Yes
Observations	16764	16764
R^2	0.421	0.623

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Effect of de-anonymized publication on compatriot score advantage

	Artistic score	Technical score
	(1)	(2)
Compatriot	0.207*	0.321***
	(0.118)	(0.089)
Compatriot × Non-JGP	0.092	-0.331**
	(0.145)	(0.134)
Compatriot × Post	-0.122	-0.283*
	(0.147)	(0.150)
Compatriot × Post × Non-JGP	0.051	0.414*
	(0.185)	(0.238)
Home event	0.600***	0.453***
	(0.076)	(0.096)
Controls for current world rank	Yes	Yes
Skater FEs	Yes	Yes
Round FEs	Yes	Yes
Observations	16106	11568
R^2	0.962	0.984

Standard errors in parentheses (clustered by event). World rank controls include the current ISU rank at the time of performance, the squared rank, as well as an indicator for being unranked. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.7: Heterogeneity of effects on within-judge consistency of subscores

	SD of artistic subscores		SD of technical subscores		
	(1)	(2)	(3)	(4)	(5)
Non-JGP	0.029***	0.026***	0.054***	0.018	0.015
	(0.005)	(0.004)	(0.016)	(0.016)	(0.017)
Post × Non-JGP	-0.020***	-0.022***	0.003	0.017	0.026
	(0.005)	(0.005)	(0.018)	(0.019)	(0.020)
Round quality × Non-JGP	0.011***	0.009***	0.034***	0.015*	0.015*
	(0.002)	(0.002)	(0.009)	(0.009)	(0.009)
Round quality × Post × Non-JGP	-0.006***	-0.005**	0.002	0.012	0.006
	(0.002)	(0.002)	(0.010)	(0.010)	(0.011)
Median score	0.005**	-0.024***	0.079***	0.095***	0.054***
	(0.002)	(0.004)	(0.013)	(0.018)	(0.018)
Median score squared	-0.001***	0.000	-0.008***	-0.013***	-0.009***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.002)
Skater FEs	—	Yes	—	Yes	Yes
Season FEs	Yes	Yes	Yes	Yes	Yes
Discipline × Segment FEs	Yes	Yes	Yes	Yes	Yes
Control group mean	0.216	0.216	1.018	1.018	1.038
Observations	150458	150458	150431	150431	108675
R^2	0.037	0.088	0.236	0.365	0.348

Standard errors in parentheses (clustered by event). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: Association between score consistency and score distance to the median judge

	Distance to the median judge					
	Artistic score			Technical score		
	(1)	(2)	(3)	(4)	(5)	(6)
SD of artistic subscores	1.219***		1.209***	0.407***		0.394***
	(0.082)		(0.082)	(0.097)		(0.098)
SD of technical subscores		0.102***	0.041		0.075*	0.056
		(0.026)	(0.025)		(0.042)	(0.042)
Constant	1.154***	1.320***	1.113***	1.698***	1.711***	1.644***
	(0.018)	(0.027)	(0.032)	(0.022)	(0.043)	(0.047)
Performance FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	102068	102041	102041	102068	102041	102041
R^2	0.128	0.122	0.128	0.173	0.173	0.173

Only observations under anonymous scoring are included, i.e. Non-JGP events before the 2016-17 season. Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: Association between score consistency and score distance to the median judge

	Artistic score	Technical score
	(1)	(2)
Artistic subscore SD	1.319***	0.572***
	(0.121)	(0.144)
Artistic subscore SD × Post	-0.255	-0.136
	(0.169)	(0.238)
Artistic subscore SD × Non-JGP	-0.055	-0.293
	(0.197)	(0.212)
Artistic subscore SD × Post × Non-JGP	0.341	0.303
	(0.255)	(0.333)
Constant	1.103***	1.817***
	(0.014)	(0.019)
Performance FEs	Yes	Yes
Observations	150458	150458
R^2	0.133	0.182

Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: Association between score consistency and the use of integer score

	Share of integer values in the PCS		
	(1)	(2)	(3)
SD of artistic subscores	0.076***		0.076***
	(0.006)		(0.006)
SD of technical subscores		0.002	-0.002
		(0.002)	(0.002)
Constant	0.463***	0.478***	0.465***
	(0.001)	(0.002)	(0.003)
Performance FEs	Yes	Yes	Yes
Observations	150458	150431	150431
R^2	0.122	0.120	0.122

Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: Effect of judge experience on within-judge consistency of scores

	SD of artistic subscores		SD of technical subscores	
	(1)	(2)	(3)	(4)
log(judge experience)	-0.0038**	-0.0027	-0.0050**	0.0046
	(0.0017)	(0.0026)	(0.0019)	(0.0032)
Constant	0.2230***	0.2197***	1.0323***	1.0046***
	(0.0047)	(0.0073)	(0.0059)	(0.0093)
Judge FEs	—	Yes	—	Yes
Performance FEs	Yes	Yes	Yes	Yes
Observations	113728	113728	113701	113701
R^2	0.198	0.381	0.850	0.861

Experience is measured as the number of competitions in which a judge has judge at, from season 2005-06 up to the previous season. Standard errors in parentheses are clustered by event. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.12: Statistics on pool of countries submitting judges to Non-GP treatment events.

Event Type	# Country	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20
European Championships	Outgoing	3	3	1	4	3	3	N.A.
	From Previous Season	N.A.	23	24	25	25	24	24
	Incoming	N.A.	4	2	4	2	3	4
	Total	26	27	26	29	27	27	28
Four Continents	Outgoing	7	11	8	7	8	9	N.A.
	From Previous Season	N.A.	20	19	20	19	20	20
	Incoming	N.A.	10	9	6	9	9	6
	Total	27	30	28	26	28	29	26
World Juniors	Outgoing	7	5	7	5	5	7	N.A.
	From Previous Season	N.A.	23	25	24	25	27	23
	Incoming	N.A.	7	6	6	7	3	6
	Total	30	30	31	30	32	30	29
World Championships	Outgoing	4	5	5	3	6	9	N.A.
	From Previous Season	N.A.	25	23	21	23	23	21
	Incoming	N.A.	3	3	5	6	7	8
	Total	29	28	26	26	29	30	29
Total	Outgoing	21	24	21	19	22	28	N.A.
	From Previous Season	N.A.	91	91	90	92	94	88
	Incoming	N.A.	24	20	21	24	22	24
	Total	112	115	111	111	116	116	112

Table A.13: Proportion of Non-JGP (Treatment) judges remaining next season.

	# Judges	% Remaining Next Season	Difference Next Season	T-test p-value
Season				
2005-06	245	0.706	0.046	0.257
2006-07	238	0.752	-0.11	0.009
2007-08	240	0.642	0.054	0.228
2008-09	207	0.696	-0.052	0.248
2009-10	230	0.643	0.019	0.682
2010-11	216	0.662	0.044	0.332
2011-12	214	0.706	0.056	0.189
2012-13	222	0.761	-0.045	0.277
2013-14	229	0.716	-0.069	0.116
2014-15	218	0.647	0.028	0.545
2015-16	215	0.674	0.049	0.268
2016-17	210	0.724	-0.085	0.06
2017-18	216	0.639	-0.043	0.366
2018-19	208	0.596	N.A.	N.A.

Table A.14: # Competitions by Non-JGP (Treatment) judges Who remain in next season.

	# Competitions Season	# Competitions Season + 1	Difference	T-test p-value
Season				
2005-06	5.734	4.965	-0.769	0.057
2006-07	5.067	5.017	-0.050	0.889
2007-08	5.286	5.143	-0.143	0.731
2008-09	5.118	5.201	0.083	0.853
2009-10	5.297	4.642	-0.655	0.124
2010-11	4.937	5.238	0.301	0.465
2011-12	5.060	4.589	-0.470	0.245
2012-13	4.219	4.941	0.722	0.085
2013-14	4.817	4.207	-0.610	0.152
2014-15	4.482	4.447	-0.035	0.935
2015-16	4.566	4.821	0.255	0.549
2016-17	4.724	5.493	0.770	0.110
2017-18	4.775	4.638	-0.138	0.774
2018-19	4.815	4.540	-0.274	0.566

Total	Not Found	Found	Percent Found
228	33	195	0.855263
218	35	183	0.839450
215	35	180	0.837209
209	31	178	0.851675
214	40	174	0.813084
208	36	172	0.826923
188	44	144	0.765957

Table A.15: Share of judges for which we could construct the nationalistic bias proxy

Total	Not Found	Found	Percent Found
228	32	196	0.859649
218	33	185	0.848624
215	35	180	0.837209
209	31	178	0.851675
214	40	174	0.813084
208	35	173	0.831731
188	44	144	0.765957

Table A.16: Share of judges for which we could construct the score accuracy proxy